NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

DETERMINING THE IMPORTANCE OF NATIONALITY ON THE OUTCOME OF BATTLES USING CLASSIFICATION TREES

by

Ali Cakan

June 2003

Thesis Advisor: Thomas W. Lucas
Second Reader: Samuel E. Buttrey

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** June 2003	**3. REPORT TYPE AND DATES COVERED** Master's Thesis	
4. TITLE AND SUBTITLE: Determining the Importance of Nationality on the Outcome of Battles Using Classification Trees			**5. FUNDING NUMBERS**
6. AUTHOR(S) Ali Cakan			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A			**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited			**12b. DISTRIBUTION CODE**

13. ABSTRACT ***(maximum 200 words)***

Throughout history people have searched for a means of predicting the outcomes of battles. Data analysis is a way of understanding the factors associated with battle outcomes. There are objective factors, such as force ratio, and subjective factors, such as leadership, that affect battles. Subjective factors are hard to determine and thus are usually avoided in models. Here, nationality is investigated as a surrogate for subjective factors. That is, we want to see how nationality is associated with battle outcomes by exploring the best available data set on historical land combat—developed by the Center for Army Analysis. We focus on four countries for which there is sufficient data: the USA, Germany, Britain and Israel. We find that these countries historically use a substantial amount of military power to defeat their enemies. In particular, the USA often has overwhelming force. Using classification tree models, with a correct classification rate of 79 percent, the results suggest that nationality was the most important factor in battles before World War I and the second most important factor during the World Wars. Force ratio was the most important factor in WWI and artillery ratio in WWII. In the years following WWII, the dominant variable has been air force ratio.

14. SUBJECT TERMS Battle Outcomes, Force Ratios, Leadership, Nationality, Historical Combat, Air Force Ratio, WWI, WWII			**15. NUMBER OF PAGES** 97
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UL

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

DETERMINING THE IMPORTANCE OF NATIONALITY ON THE OUTCOME OF BATTLES USING CLASSIFICATION TREES

Ali Cakan
First Lieutenant, Turkish Army
B.S., Turkish Military Academy, 1998

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS ANALYSIS

from the

NAVAL POSTGRADUATE SCHOOL
June 2003

Author: Ali Cakan

Approved by: Thomas W. Lucas
Thesis Advisor

Samuel E. Buttrey
Second Reader

Jim Eagle
Chairman, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Throughout history people have searched for a means of predicting the outcomes of battles. Data analysis is a way of understanding the factors associated with battle outcomes. There are objective factors, such as force ratio, and subjective factors, such as leadership, that affect battles. Subjective factors are hard to determine and thus are usually avoided in models. Here, nationality is investigated as a surrogate for subjective factors. That is, we want to see how nationality is associated with battle outcomes by exploring the best available data set on historical land combat—developed by the Center for Army Analysis. We focus on four countries for which there is sufficient data: the USA, Germany, Britain and Israel. We find that these countries historically use a substantial amount of military power to defeat their enemies. In particular, the USA often has overwhelming force. Using classification tree models, with a correct classification rate of 79 percent, the results suggest that nationality was the most important factor in battles before World War I and the second most important factor during the World Wars. Force ratio was the most important factor in WWI and artillery ratio in WWII. In the years following WWII, the dominant variable has been air force ratio.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I am grateful to my beloved country for giving me everything that I have.

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Throughout history, predicting the outcome of a battle before it starts has been a main concern of soldiers, historians and analysts. Different tools have been used to make predictions. Two of the most important and commonly used are simulation and data analysis.

Built on mathematical models, such as Lanchester equations, simulations, especially with the advancements in computer technology, are becoming increasingly important. Recent developments in computer technologies and new algorithms have made simulations very capable and reliable, but there still are pitfalls. For example, it is difficult to model intangibles such as leadership and training, and these factors can be just as important as a soldier's weapon.

Another tool is data analysis. It is widely used and has been producing quite satisfactory results. Moreover, unlike simulations, it is possible to use intangibles in data analysis models. In this work, we use data analysis. Our interest area is nationality factors. In other words, do different nations have different characteristics that affect the outcome of a battle? If there are nationality factors, what are they? Do they change over time? Can we use them to predict the outcome of a potential battle?

In our analyses, we used the CDBG90 data set, developed for the Center for Army Analysis (CAA). This is the best data set available on historical land combat. This data set was first prepared by the Historical Evaluation and Research Organization (HERO) in 1983, and we are using the version with the latest updates. The CDBG90 includes 657 battles from 1600 to the end of the 20th century. There are up to 152 attributes listed for each battle.

Numerous people have worked with this data set, including some NPS Masters' students. These researchers looked at different aspects of warfare and tried to answer different questions. The first analysis was done by CAA, under the Combat History Analysis Effort (CHASE) beginning in 1984. Afterwards, Dupuy [Ref. 2] tried to model

warfare without using advanced analysis techniques and formed the Quantified Judgment Model. Hartley built his Oak Ridge Spreadsheet Battle Model, which allows the user to predict the outcome of a potential conflict using an Excel spreadsheet [Ref. 1]. Yigit looked at the famous rule of thumb that an attacker with greater than a 3:1 Force Ratio wins, and questions such as "How successful are the attackers? Do attackers suffer more casualties?"[Ref. 3]. Coban [Ref. 4] used classification trees to build a model which predicts the outcome of a potential battle.

Among the works mentioned above, Hartley's claimed that nationality factors should have an important role in modeling warfare. There is another work on this subject which is of interest to us. Prior to the Gulf War, a British analyst, David Rowland, made accurate predictions about the results of the war, relying heavily on nationality factors [Ref. 5].

To do the analyses, we divide the data set into four subsets with respect to the time, because the nature of warfare changes as time evolves and battles in these time periods have similar characteristics. The first subset, battles before World War I (WWI) covers the battles from 1600 to the beginning of WWI. The second subset is the battles in WWI, the third subset is the battles in WWII, and the last subset is the battles after WWII. We also focus on four countries, the USA, Germany, Britain and Israel, because more data are available in the data set on these countries than the others.

Our first analysis is done with the objective variables, namely force ratio, tank ratio, artillery ratio, air force ratio and cavalry ratio. These come from *hard data*. That is, the values for these variables can be actually collected from the battlefield. We use boxplots to show the data structure and Wilcoxon's rank sum test to compare different hypotheses relating to the objective variables. We find that the USA has usually accumulated great power on the battlefield. Especially in WWII, the air force and tank ratio of the USA is overwhelming, almost incomparable to those of their enemies. We see either little or no difference between Germany and Britain, and also, we usually did not see a statistically significant difference between the ratios of countries when they won or lost. Among the countries, Israel, has the smallest figures for all objective variables, except for air force ratio.

We also examine the relative variables. These variables, such as initiative, leadership, and training, come from *soft data*, i.e., the values of them are decided by the judgment of historians. Therefore, they are subjective and are usually avoided in models. Our analyses showed that the data set does not have useful information for these variables. For most of the battles, neither side was deemed to have an advantage with respect to relative variables and the countries had similar patterns. Only Israel has a different pattern than other countries. For training, leadership and combat effectiveness advantage, they have an obvious advantage over their opponents in the battles they fought.

We used classification trees in our final analyses to see if the nationality factors were important. Tree-based modeling is an exploratory technique for uncovering structure in data and is useful for summarizing large multivariate datasets. [Ref. 7] Trees do not need distributional assumptions, and interactions between variables are automatically included in the tree structure. In addition, they are robust to outlying data. One of the advantages of tree-based models is that they are easy to read. There are oval (non terminal or split) and rectangular (terminal) nodes. Each node contains the predicted outcome and the distribution to the child nodes. The split criterion is shown on each branch.

The first model consists of the battles prior to WWI (Figure 1). In this period, tree models show that nationality was the most important variable, that is, the first split criterion is nationality. The second important variable is force ratio. This model explains 76 percent of the battles, that is, the model classifies 76 percent of the battles correctly. The second model, the battles in WWI, showed that nationality is the second most important variable after force ratio. The second model explained 79 percent of the battles. The third model was built with the information on the battles in WWII. Nationality is the second important variable after artillery ratio. This model also explained 79 percent of the battles. The last model, the battles after WWII, consists of the battles in which Israel participated. The only variable that appears in this model is the air force ratio. To evaluate the importance of nationality in our tree models, we fit the models with and without nationality factors and compared the misclassification rates of the two.

Nationality factors improved the accuracy of the model for the battles prior to WWI, but the improvement was insignificant for the models for WWI and WWII and did not appear in the last model.

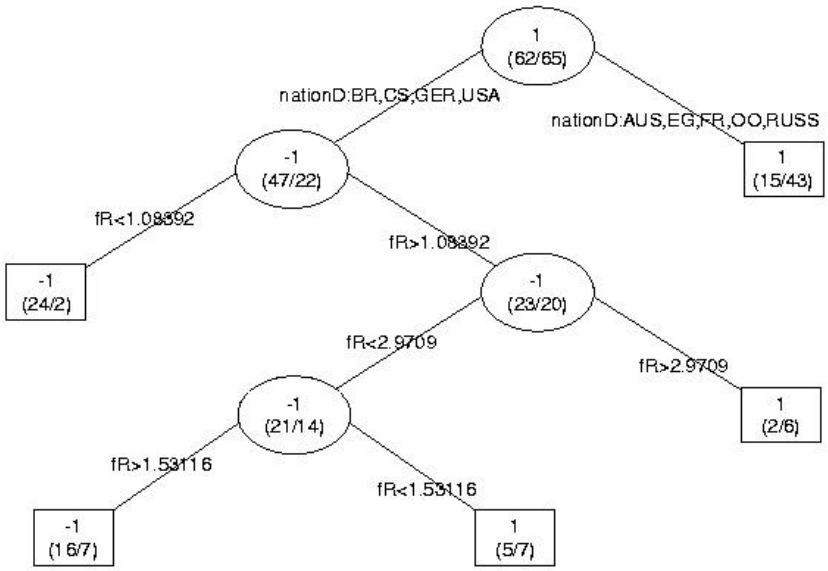


Figure 1. Tree Model for the Battles Before WWI.

> Figure 1 is our first model which includes the battles before World War I. The most important factor is nationality. If the defender is from one of the following, the USA, Britain, the Confederate States or Germany, the model suggests predicting a win for the defender. If we were to predict an outcome of a hypothetical battle in this period, we could predict the result only by looking at the nationality of the countries and we would be correct 71 percent of the time. After nationality, the single most important variable is force ratio.

Coban [Ref. 4] found that relative variables were the most important factors before WWI. Our model for that period, without using any relative variables, explained 76 percent of the battles versus Coban's [Ref. 4] 79 percent. This shows that we can replace the relative variables with the nationality variable, and still have a pretty good

model, at least for this data. This is totally objective, because nationality is known before the war starts, whereas the relative variable values are very difficult to determine, and vary from analyst to analyst. The models for the other periods did not show the nationality variables to be the most important factor. However, combining the results from all other analyses with the results of our classification trees, we conclude that having sufficient military power on the battlefield is a nationality factor for all four countries.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. INTRODUCTION

When one reads about history, it is mostly the history of wars. It is not too far of a stretch to say that wars shaped our history, and will continue to be one of the most important phenomena shaping the future of the world. Having this much importance, a lot of effort has been, and is being, devoted to exploring "the art of war". One of the main areas of interest has always been predicting the outcome of a battle before the first bullet flies. Related to, and probably more important than, this question is "what relates to winning?" As discussed below, many researchers, using different tools, have tried to answer this question.

Simulation is one of the tools used to make predictions about potential battles. Simulations are often built on mathematical models, such as Lanchester equations [Ref. 4]. In the past, capabilities of simulations were somewhat limited, but with improvements in computer technology, much more capable simulations are available today. In the end, though, simulations are simplifications of combat, and it has proven difficult to model intangible factors, such as like leadership, morale, training etc. [Refs. 1, 2, 3], which according to many other studies, greatly affect combat outcomes.

Another tool that analysts use to understand the nature of warfare is data analysis. The main challenge with data analysis is finding reliable, useful data. Furthermore, the data need to be detailed and large enough to find reliable answers. To some extent, we also have this problem, but the data set used in this research is considered to be the best data set available on historical land battles. In this work, the CDB90G data set is used. It is an updated version of the data set consisting of historical data prepared by the Historical Evaluation and Research Organization (HERO) in 1983, which includes battles from 1600 through the Arab-Israeli wars towards the end of the 20^{th} century.

In 1983, the U.S. Concepts Analysis Agency (CAA) contracted the Historical Evaluation and Research Organization (HERO) to build a data set of historical combat comprising 601 battles. CDB90G, the updated version, consists of 657 battles. There are up to 152 attributes listed for each battle.

Numerous people have worked with this data set, including some NPS Master's students. These researchers looked at different aspects of warfare and tried to answer different questions. The first analysis was done by CAA, under the Combat History Analysis Effort (CHASE) beginning in 1984. Afterwards, Dupuy [Ref. 2] tried to model warfare without using advanced analysis techniques and formed the Quantified Judgment Model. Hartley built his Oak Ridge Spreadsheet Battle Model, which allows the user to predict the outcome of a potential conflict using an Excel spreadsheet [Ref. 1]. Yigit looked at the famous rule of thumb that an attacker with greater than a 3:1 force ratio wins, and questions such as "How successful are the attackers? Do attackers suffer more casualties?"[Ref. 3]. Coban [Ref. 4] used classification trees to build a model which predicts the outcome of a potential battle.

Coban's work in particular is interesting since it uses the relatively new data analysis method of classification trees to model combat. Tree based models have certain advantages over traditional linear models. They are usually easier to discuss and interpret than linear models, and the treatment of missing values (NAs) is more satisfactory with tree based models than linear-based models [Ref. 9:p. 378]. Being easier to understand, the models built can easily be used by people not very knowledgeable in the subject.

Among the works mentioned above, Hartley's claimed that nationality factors should have an important role in modeling warfare. Prior to his research, the other works discussed did not emphasize the importance of nationality factors as much as did Hartley. There is another example of the importance of the nationality factors from a British analyst, David Rowland. Prior to the Gulf War, among all the predictions on the outcome of the operation, his was reportedly the most accurate [Ref. 5]. Rowland used nationality factors as his main variable. Before the campaign started, while most analysts were estimating that the battle would last for months and cost thousands of allied lives [Ref. 6], he predicted that it would be easy, and came up with figures close to what happened. These two works, especially the second one, motivated this thesis.

The hypothesis that this thesis investigates is that a phenomenon called "Nationality Factor" exists; that is, every nation has its own characteristics, which also affects its military. For example, people think that the Germans have a long military

tradition and are good fighters. Indeed, Dupuy estimates that one German soldier had more combat effectiveness than two Soviet soldiers [Ref. 2]. Japan has its own fighting class, the samurai, who have hundreds of years of tradition, which makes the country's military unique and different from other countries. In Turkey, being a soldier is special. It is said that "Every Turk is born as a soldier," and it is a great honor to die in a battle for a person and his family. Many more examples can be found. It is an undeniable fact that there is more to winning than having more weapons or superior tactics or perhaps even better training. It is interesting to look at battles where the side with apparently less power was the victor, not only once, but many times. The recent Arab-Israeli Wars are a clear example where an outgunned side (Israel) repeatedly won.

The purpose of this thesis is to search for the presence of a "Nationality Factor" and find its effects, if any using the CDB90G data set.

B. BACKGROUND

1. Trevor Dupuy

A retired U.S. Army colonel, Trevor N. Dupuy, founded the Historical Evaluation and Research Organization (HERO), which constructed most of the data set used in this thesis. After finishing the data set, he did an analysis using the very same data set. The main product was the Quantified Judgment Model (QJM). The formulas in the QJM model are only a little more complicated than basic math. A main feature of QJM is the use of OLI values. OLI stands for Operational Lethality Index, which is a weapon's maximum effect under ideal conditions [Ref. 2:p. 30]. The effects of the battlefield, i.e., the changes from "ideal" conditions, are represented by different variables. The combat power computation is built upon the corrected (the effects of variables included) OLI factors and as an end product, the outcome value R (for result), is calculated for both sides. If R_f—R_e (R_f: Result friendly, R_e: Result enemy) is positive, the model predicts that the friendly side wins, and vice versa. Analysis is done to calculate the variable values and effects. Nationality factors are not extensively used. [Ref. 2]

2. Dean Hartley

In his book, “Predicting Combat Effects”, Hartley analyzed the original HERO dataset to determine whether there are any consistent formulae for predicting combat effects. The results proved to be positive and were incorporated in a spreadsheet model that predicts battle outcomes; including attrition, duration, advance, and victory.

Attrition, at the gross level, is determined to follow neither the Lanchester Square Law nor the Lanchester Linear Law. Instead, it follows a law between the Linear Law and the Logarithmic Law. See Equation (1)

$$\frac{d}{dt}E = - E^{0.75} F^{0.40}$$
$$\frac{d}{dt}F = - F^{0.75} E^{0.40} \quad (1)$$

where

E = enemy manpower,

F = friendly manpower,

t = time.

More extensively than the other works, Hartley used nationality factors as one of the more important variables and it appears in many of his computations. For example, the predicted log duration of a battle is: p. 95:

PFLDURA2=.31 + .24*AIRPL + .0000083*STARDAT- 157*TEMP + .00043*RXODP - .00047*ABAIYART + .000054*LWIDYART+ .91*ATVAL + .96*DEVAL ; where ATVAL:

if ATTACKER =	"Arabs"	then ATVAL =	0.5
if ATTACKER =	"Austria"	then ATVAL =	0.2
if ATTACKER =	"England"	then ATVAL =	0.2
if ATTACKER =	"European"	then ATVAL =	0.4
if ATTACKER =	"France"	then ATVAL =	0.0
if ATTACKER =	"Germany"	then ATVAL =	0.0
if ATTACKER =	"Israel"	then ATVAL =	0.3
if ATTACKER =	"Italy"	then ATVAL =	0.8
if ATTACKER =	"Japan"	then ATVAL =	0.0
if ATTACKER =	"Other"	then ATVAL =	-0.1
if ATTACKER =	"Russia"	then ATVAL =	0.2
if ATTACKER =	"USA"	then ATVAL =	0.0

By the same token, nationalities have different constants in almost every calculation. Thus, they greatly affect the end result.

Despite the extensive usage and its benefits, using nationality factors is still considered suspect. The reason is the different nature of national identity. It is not fixed and certainly does change over time. Once it was the Romans who ruled the world, France enjoyed military superiority from the time of Napoleon to the Franco-Russian War, which they lost. In another example, the Ottomans were the main power for centuries and then they became "the sick man of Europe." [Ref. 1]

3. Faruk Yigit

Yigit [Ref. 3] explored CAA's revised version of the HERO database, the CDB90FT. This dataset consists of 660 battles and engagements with up to 140 different attributes on each. Yigit analyzed the 3-1 force ratio rule of thumb, the dispersion rate, and the daily casualty rate. He divided the data into chronological subsets and analyzed each subset. He concluded that force ratio was a reasonable predictor of outcomes. For example, a force ratio of 3 to 1 or greater leads an attacker to victory 68 percent of the time. Some of his other findings are that greater dispersion of combat troops is a reason for the decrease in casualties despite an increase in weapon lethality, and casualty rates of the attacker are almost always lower than those of the defender.

4. Muzaffer Coban

Coban [Ref. 4], using the latest version of the data set, CDB90G, used classification trees to build models that predict the outcomes of potential battles. Tree-based methods may be unfamiliar to some analysts, although many researchers like them since they present an attractive way to express knowledge and aid in decision making [Ref. 9:p. 251]. Coban looked at pre-selected variables, which he thought had more of an effect on the outcome of battle. The pre-selected variables were analyzed to show descriptive statistics and conditional plots. The pre-selected variables were:

- **Objective variables:** force ratio, tank ratio, artillery ratio, cavalry ratio, the attacker's primary tactical scheme, and the defender's primary defensive posture.

- **Relative variables:** relative surprise, relative air superiority in the theater, relative combat effectiveness, relative leadership advantage, relative training advantage, relative morale advantage, relative logistics advantage, relative momentum advantage, relative intelligence advantage, relative technology advantage, relative initiative advantage.
- **Terrain and weather variables:** three terrain factors and five weather factors.

The descriptive statistics and conditional plots revealed the association of the variables with the outcome of battles. The descriptive statistics revealed that the objective variables are not highly correlated with victory. Some of the relative variables, such as leadership, have a strong relationship with victory. However, relative variables are subjective and based on historical judgment.

Using these variables, three tree-based models were considered. Model 1, with only the objective variables, resulted in high misclassification rates. This result was parallel to the findings with descriptive statistics, which was that objective variables alone are not sufficient to classify battle outcomes. Model 2, with both objective and relative variables had relatively low misclassification rates. Model 3 used terrain and weather variables, as well as the objective and relative variables. However, the resulting classification trees did not include the terrain and weather variables, and the misclassification rates were no better then those of Model 2.

Coban conducted another analysis to understand the historical trends in battles. Multiple classification trees were built by using the objective and relative variables with training test sizes of 125. Each classification tree was built with a training set size of 125 and the battle after the 125 battles in the data set was predicted. Then, another classification tree was built with the next 125 battles, with an overlap of 124 battles. At the end, 658-125=533 classification trees were built and 533 predictions made. This analysis revealed some important results. First, the importance of variables has changed throughout history. Second, the misclassification rates show that past battles failed to predict the battles of World War II, in which new tactics and weapons were introduced to fighters [Ref. 4].

In his thesis, Coban concluded that:

> The predictions of battle outcomes using classification trees revealed as high as 79 percent correct (clear-cut outcomes). This result is satisfying when the role of luck in battles and hard to quantify factors are considered. [Ref. 4]

This is the most interesting part, hard to quantify variables, which resulted in being the topic of this thesis.

It is always a challenge to work with intangibles. How can you measure things such as leadership or morale, especially before a conflict? Can nationality factors be a surrogate for these?

In the CDB90G data set, there are values for the intangible variables as well, but since the purpose is to predict the outcome of a war, we need data before the war, not afterwards. However, I have a different point of view. I hypothesize that nations have their own characteristics, which are force multipliers. A good thing about this particular variable is that although it is a soft factor, the nationality factor, unlike other soft factors, is “objective”. The purpose of this work is to ascertain if nationality factors correlate with the outcome of battles above and beyond other variables. What are the nationality factors and can we really talk about them? If the answer is yes, do they change over time? Can we come up with a reasonable method to use nationality factors in predicting the outcome of a battle?

THIS PAGE INTENTIONALLY LEFT BLANK

II. SUMMARY STATISTICS

A. INTRODUCTION

This section explores and summarizes the data set using simple analysis techniques. Our purpose is to establish a good fundamental understanding of the data set before actually doing analysis with classification trees.

To address the purpose of the thesis, that is to determine the effect of nationality on the outcome of a potential battle, the data are analyzed with respect to different nationalities. In order to do this, we use different subsets of the data set with respect to different nationalities. Variable "nationA", the nationality of the attacking force, is used as the classifying variable. In addition, "nationD", the nationality of the defending force, is also used when necessary.

One of the questions we want to answer is whether nationality factors change over time. To address this particular aspect, following Coban [Ref. 4], the data set is divided into six different time periods. These time periods reflect important changes in history. In each period, war was conducted differently than in the others, in that new technologies or new tactics were used. The battles within each period have similar properties. The first division is made at 1755, and therefore, the first period is 1600 to 1755. The Thirty Years' War falls within the first period. 1756 marked the beginning of the 7 Years' War, which was the largest of the pre-Napoleonic Wars in the data set. The second period is from 1756 to 1814, and includes the 7 Years' War and the Napoleonic Wars. This was the period of great European powers, extensive usage of black powder and big sailing ships. 1815 marks the fall of Napoleon, and the beginning of a new era. In this period, from 1815 to 1914, a big portion of the data is from the American Civil War. This period ends in 1914, the beginning of World War I (WWI). 1914 to 1939 comprise the next period. This is mostly WWI, in which warfare changed in revolutionary ways, as many new technologies, such as tanks, airplanes and chemical warfare were used. The next period, from 1939 to 1945, has data from World War II (WWII). This is the most important subset because it has more data on the nations that we are interested in than the other subsets and the way battles were fought more closely resembles today's concepts.

Another advantage with this period is that the data is more reliable because record keeping was much better than before WWII. The last period is from after WWII to the present.

The number of battles of different countries in different periods in the data set is shown in the tables below. For acronyms, see Appendix C.

RowNames	1600+ thru 1755	1755+ thru 1814	1814+ thru 1913	1913+ thru 1939	1939+ thru 1945	1945+ thru 2000	total
AUS	4	11	0	7	0	0	22
BR	1	12	7	24	26	0	70
CS	0	0	22	0	0	0	22
ENG	10	0	0	0	0	0	10
FR	14	28	5	12	1	0	60
GER	0	0	8	26	45	0	79
IS	0	0	0	0	0	49	49
OO	29	20	26	17	3	23	118
PR	4	9	1	0	0	0	14
RUSS	0	0	3	4	0	0	7
SOV	0	0	0	3	24	0	27
USA	0	2	35	40	94	8	179
TOTAL	62	82	107	133	193	80	657

Table 1. Battles Per Period, Attacker.

RowNames	1600+ thru 1755	1755+ thru 1814	1814+ thru 1913	1913+ thru 1939	1939+ thru 1945	1945+ thru 2000	total
AUS	0	14	3	11	0	0	28
BR	2	10	4	7	11	0	34
CS	0	0	27	0	0	0	27
EG	0	0	1	0	0	26	27
FR	6	24	8	10	3	0	51
GER	0	0	1	69	110	0	180
IMP	12	0	0	0	0	0	12
IS	0	0	0	0	0	20	20
JAP	0	0	2	3	31	0	36
OO	37	29	27	5	4	17	119
RUSS	1	4	5	10	0	0	20
SOV	0	0	0	5	6	0	11
SYR	0	0	0	0	0	14	14
TU	4	0	5	11	0	0	20
USA	0	1	24	2	28	3	58
TOTAL	62	82	107	133	193	80	657

Table 2. Battles Per Period, Defender.

Tables 1 and 2 show the number of battles in which countries were involved during different periods. The first table contains the numbers for when the country was an attacker (nationA), the second when it was the defender (nationD). The countries that will be analyzed are highlighted. As an example, the USA has 179 attacks, 94 of them in WWII, and Germany has 180 defends, 110 in WWII.

Tables 1 and 2 reveal a problem. Although the data set includes 657 battles, the number of battles decreases dramatically when divided into subsets, making analysis difficult. To overcome this problem, the following is done.

1600 to 1914 is considered as a single period. This follows Coban [Ref. 4], who considered the battles prior to WW I as a group. He showed that intangibles are the most important factors in this period.

The names of the following countries are combined. BR and ENG, PR and GER, SOV and RUSS. CS (Confederate States) and the USA are not combined, because the author considered them different countries since the battles of CS were against the USA. This is also in line with the way Hartley analyzed the data [Ref. 1]. Also, again for the same purpose, the focus will be on four Nations: the USA, Germany, Britain, and Israel. The new tables of battles per period for the four countries we will analyze are in Tables 3 and 4.

	1600+ thru 1913	1913+ thru 1939	1939+ thru 1945	1945+ thru 2000	total
USA	37	40	94	8	179
BR	30	24	26	0	80
GER	22	26	45	0	93
IS	0	0	0	49	49

Table 3. Battles Per Period, Attacker.

	1600+ thru 1913	1913+ thru 1939	1939+ thru 1945	1945+ thru 2000	total
USA	25	2	28	3	58
BR	16	7	11	0	34
GER	1	69	110	0	180
IS	0	0	0	20	20

Table 4. Battles Per Period, Defender.

B. DESCRIPTIVE STATISTICS

In this section, the important variables of the CDB90G data set will be analyzed. Fifteen different variables are considered as potentially important, that is, potentially affecting the outcome of the battle. This decision follows the selections made by Coban [Ref. 4] and Hartley [Ref. 1], and also reflects the author's military judgment. The variables are divided into two subsets of "objective" and "relative" variables. Objective variables are those whose values can be collected from the battleground or from "hard" data. They are force ratio, artillery ratio, air force ratio, cavalry ratio, and tank ratio. These variables can be known before the confrontation and can be agreed upon by different people. While the accuracy of this data is suspect [Ref. 1], they are based on numbers, so they have the same meaning for everybody. As an example, one tank is one tank for all analysts. Therefore, these variables are called objective variables. On the other hand, relative variables, leadership, training, combat effectiveness, are totally subjective; the values being based on the judgment of military historians. Unlike the case with objective variables, it is extremely difficult to decide the values of these before the battle, and differences between different people's figures are almost guaranteed. Therefore, they are called "soft" data, and are almost universally avoided in models [Ref. 1]. We will not be an exception.

For our purposes, then, objective variables are much more important than relative variables. There are other works, Hartley [Ref. 1] and Coban [Ref. 4], which used relative variables in their models. We will follow a different method. All variables will be analyzed in this section to reveal characteristics of different nations, but after this section, the relative variables will not be analyzed again. Instead, we will try to replace all the relative variables with just one variable: Nationality.

1. Treatment of the Data

In the data set, some relative variables, relative combat effectiveness, leadership, training, morale, logistics, momentum, intelligence, technology, and initiative, have values ranging from "–4" to "+4." A value of "–4" shows that the variable very strongly favors the defender, while "+4" shows that the variable very strongly favors the attacker. A level of "0" favors neither side. The variable surprise is given in a scale between "–2" and "+2." Again, negative values favor the defender and positive favor the attacker.

However, it is very difficult to scale these qualities in this much detail. Therefore, following Coban's methodology, we will give those variables only 3 values: "A" for an advantage to the attacker, "D" for an advantage to the defender, and "O" for no advantage to either side. [Ref. 4]

Since Coban also used the same data analysis method, classification trees, in his models, we will try to follow him when possible, and try to compare our findings with his. As in his analysis, weapons effects are expressed as ratios. In some battles, the attackers had no weapons of a particular type. This makes the ratio zero, which gives no information about the number of the defender's weapons. In some other cases, the defender had no weapons and that makes the ratio infinity. Adding a constant to both sides avoids these two pitfalls. Therefore, in finding ratios, one is added to each side's strength. When neither side had a particular weapon system, e.g. tanks, a missing value indicator is assigned to the ratio variable. [Ref. 4]

Descriptive statistics will help understand the properties of objective and relative variables and nationalities. Tables, boxplots, barplots and histograms are used when informative.

2. Response Variable

a. Battle Outcome: "WINA"

The outcome of the battle is expressed in variable "WINA". A value of "1" represents an attacker win, "-1" means that the attacker did not win. Either the attacker lost or the historians judged that the battle was a draw.

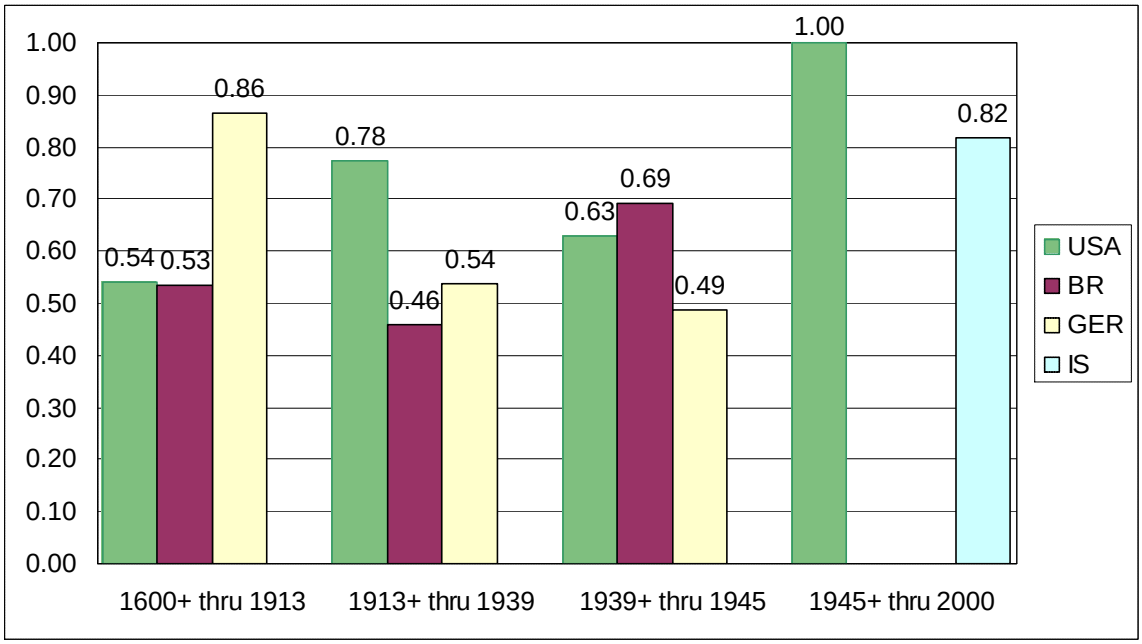


Figure 2. Proportion of Battles Won By Attacker.

Figure 2 shows the ratio of the battles won to all of the battles fought (Battles Won/Battles Fought | Attack). Israel came into existence only for the last period, and we do not have any data from Britain or Germany after WWII. For this reason, there are some missing bars. Before WWI, the Germans won a large portion of the battles when they were attackers, and this ratio consistently fell in later periods. The USA's hundred percent refers to battles in the Korean War.

3. Objective Variables

As mentioned in the introduction, the objective variables to be analyzed include force ratio, artillery ratio, tank ratio, cavalry ratio and air force ratio. All of the countries are analyzed when they were attacking. Although analyses with nation defending were done as well, they are not presented here because the results were not useful.

a. Force Ratio

The basic formula for force ratio is:

$$FR=A/D,$$

where

A is the total strength of the attacker in manpower and

D is the total strength of the defender in manpower.

The strength refers to only the combatants, and troops on either side are assumed to be identical. The following is a boxplot of the force ratios.

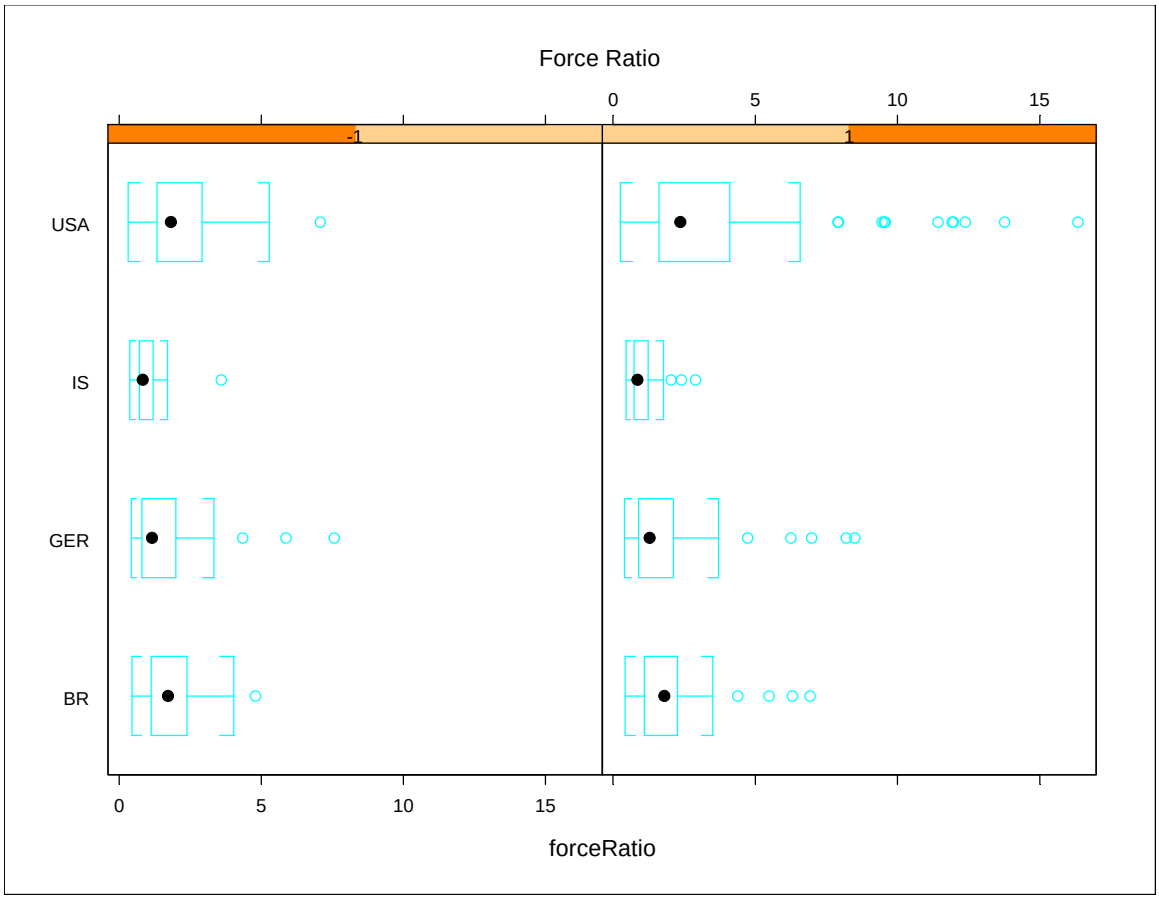


Figure 3. Force Ratios of Attacking Countries.

The first boxplot will be explained in detail. This plot is drawn by the "`bwplot()`" command in S-Plus version 2000 [Ref. 13]. This function enables us to draw boxplots for multiple variables, in this case attackers, on one chart. The force ratios are on the *X* axis, and the names of the countries are on the *Y* axis. The graphic is divided into two sections by a vertical line. Above the section on the left reads the number "-1", which refers to the value of the "Outcome Variable", "WINA". As explained in the respective section, -1 means that the attacker did not win. So, while the force ratio boxplots of the battles that the attacker lost are on the left, the ones that the attacker won are on the right.

The rectangle-like shapes in the plot are called “boxplots”. In recent years, boxplots have successfully been used to describe the prominent features of data sets. These features include center, spread, the extent and nature of any departure from symmetry, and identification of outliers [Ref. 8]. The point in the center shows the median. The width of the rectangle is an indicator of variability, the wider the rectangle, the more the variability is, and the width of the rectangle is called the fourth spread, fs. Data between the first and the third quartiles (the middle 50 percent) fall in this rectangle. The left end of the rectangle (lower fourth) is the median of the smallest n/2 observations, and the right end (upper fourth) is the median of the largest n/2 observations. The whiskers on both sides have the smallest and the biggest observations, unless there are outliers. Any observation farther than 1.5 fs from the closest fourth is an outlier and represented as a small circle.

Force ratio is universally considered to be an important factor in battle outcomes. When Figure 3 is examined, it can be seen that there are differences between countries. The first difference is their force ratios. The next table has the average force ratios of the countries in which they won and lost while they were attacking.

Country	Lose	Win
BR	1.93	2.07
GER	1.68	1.93
IS	1.20	1.05
USA	2.22	3.39

Table 5. Force Ratio Averages.

The USA has the highest average force ratio. Israel has the higher force ratio in the battles it lost than the battles it won.

As it can be seen from Table 5, the USA has a bigger force ratio than the others. The USA has three times more force ratio than Israel, which always has a smaller force ratio than the other three countries. Israel also has less variability. The boxplots in Figure 3 are almost symmetric, that is, the distribution of force ratios for a particular country when they won and lost, are almost identical, except for the USA. Normally, the force ratio in the battles won is expected to be higher than in the ones lost. Wilcoxon's rank-sum test is used to see whether the median force ratio of Germany and Britain in the battles they won is greater than the ones they lost. Since we are using the whole

population, the answer to this question is intuitive and does not require any statistics. It is only necessary to compare the medians. Wilcoxon's rank sum test is being used to see if the differences in median force ratio when winning and losing for Germany and Britain are indistinguishable from what would be obtained by random samples from the same distribution.

$$H_0 : m_1 - m_2 = 0$$
$$H_a : m_1 - m_2 > 0$$

where,

m_1 = Median force ratio when attacking and winning
m_2 = Median force ratio when attacking and losing

The Wilcoxon's rank-sum test reveals a p-value of 0.4845 for Britain and 0.183 for Germany. Both of these values strongly suggest that, at a five percent significance level, for both countries, there is no evidence to reject the null hypothesis. That is, the medians are the same. In other words, neither of the countries had a significantly higher force ratio for the battles they won than the battles they lost. When the medians of Germany and Britain's force ratios when they won are compared, the p-value is 0.071. This suggests a difference in their force ratios but the hypothesis that the medians of these two countries when attacking and winning are the same cannot be rejected at the 0.05 significance level.

The USA definitely has a bigger ratio and more spread when they won attacking as compared to when they lost defending. The final interesting feature is that, not only did the USA have a higher force ratio, but it also has many outliers. The numerical dominance of the USA on the battlefield and its effects will be discussed later.

b. Artillery Ratio: "arty"

arty = A_A / A_D

where

A_A = Number of artillery tubes of the attacker and

A_D = Number of artillery tubes of the defender.

This is the only variable which is present in all periods. Prior to WWI, the artillery ratio varied a lot [App. B.B.]. After the start of the 20th century, artillery was used extensively.

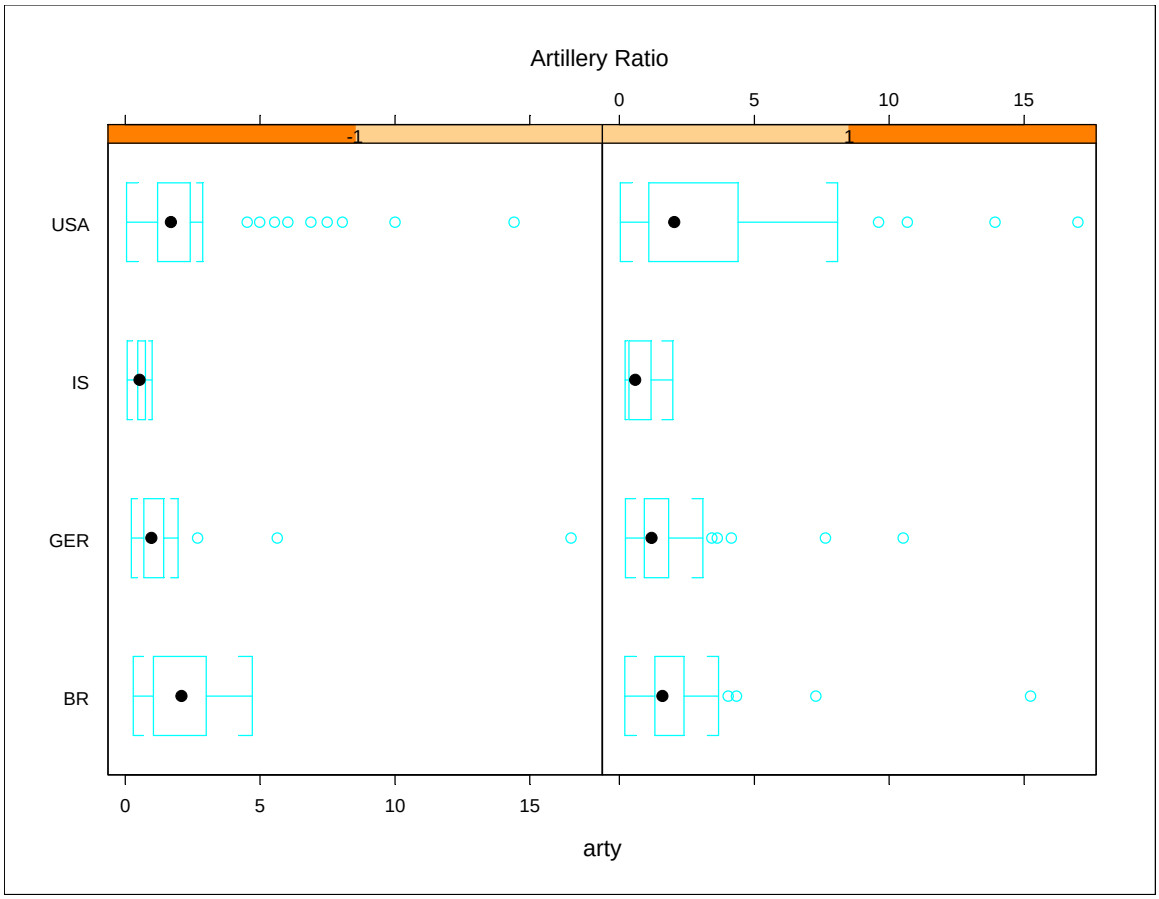


Figure 4. Artillery Ratio, Entire Data Set.

Figure 4 shows the boxplots of artillery ratios in the entire dataset. The battles with an artillery ratio more than 20 are not included for the sake of interpretability. One point is worth mentioning. During WWII, the USA has an average artillery ratio of 8.56. This is very much affected by a very big advantage, 20.18 in 1944.

Figure 4 suggests that, like force ratio, there are differences in artillery ratios between different countries as well. The USA again had a very large advantage compared to other countries, and more so in the battles they won. Like force ratio, many outliers can be seen in the USA's boxplots. Israel has the smallest advantage compared to other forces. Britain's artillery ratio looks higher in the battles they won than the battles

they lost. Germany's artillery ratio when they won and when they lost look similar. Again, as we did with the force ratios, the Wilcoxon's rank-sum test is used to test whether the artillery ratio when winning is greater than losing for Britain and Germany:

The test reveals a p-value of 0.03161 for Germany. This suggests that, at a five percent significance level, the artillery ratio when winning is higher than losing, as expected.

For Britain, the p-value is 0.5193 strongly suggests that the median artillery ratios when winning and losing are indistinguishable.

c. Close Air Support Ratio: "fly"

$fly = F_A / F_D$

where

F_A = Number of close air support sorties of the attacker and

F_D = Number of close air support sorties of the defender.

Close air support is very important in today's warfare. After armies around the world began to use them in combat, airplanes became one of the most important factors. Coban [Ref. 4] found that it is the most important variable in wars after WWI. Today, the first Gulf War and operations in Serbia have proven that an air force is a dominant factor in defining the outcome of a battle.

Although airplanes were used in WWI, there is so little data in the data set that we decided to start with WWII, which is the first war in the data set in which air forces played a major role in the outcome of battles.

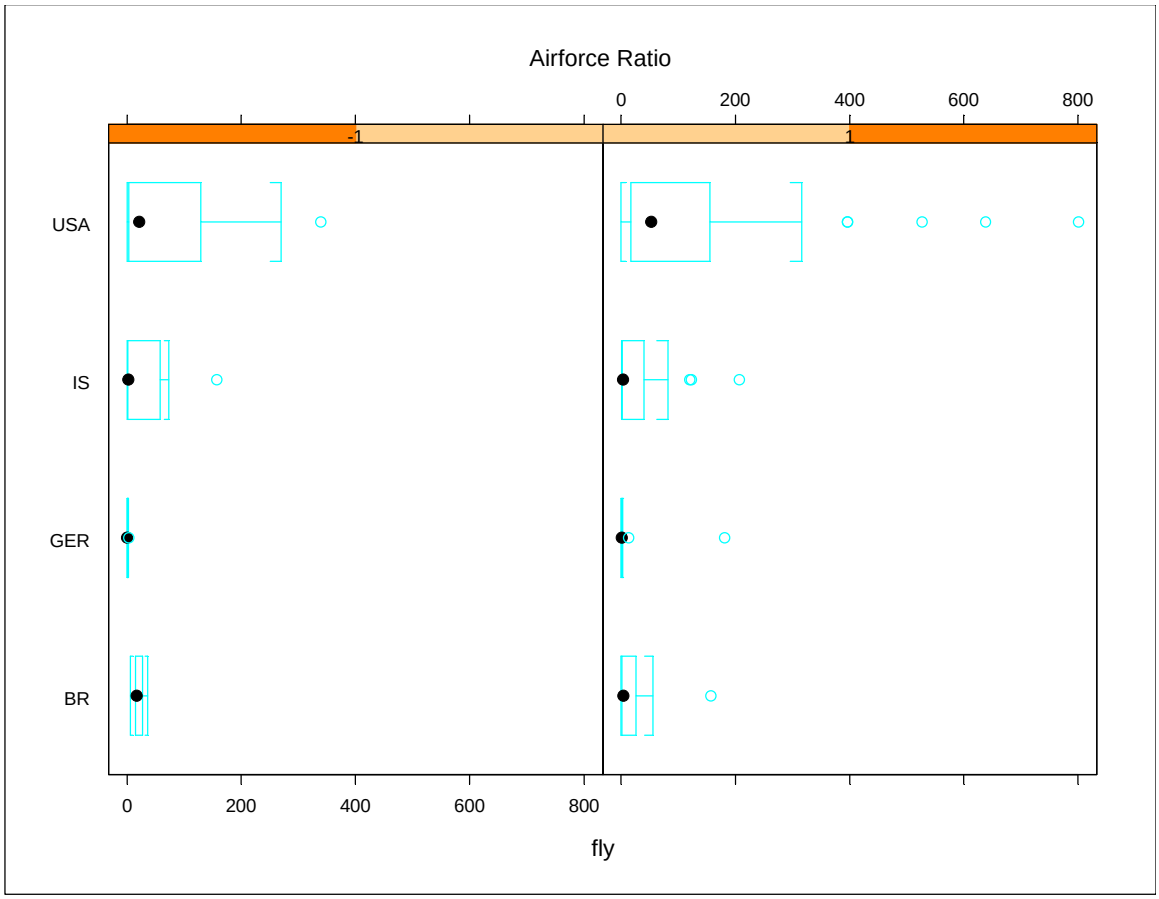


Figure 5. Air Force Ratio, All Dataset.

Figure 5 contains all battles post-WWI. It is very similar to the plot drawn with the data from WWII [App. B.C.], the difference being data from the Arab Israeli wars and the Korean War.

The USA used airplanes much more than other countries. The graph is affected greatly by the very high figures of the USA. Therefore, it is difficult to read other countries' boxplots. Unlike artillery ratio, we did not worry about truncating the data at a particular point this time because the difference is very large. The USA's overwhelming dominance with respect to the air force is an undeniable fact.

Among other countries, Israel used its air force more than Germany or England. Although countries other than Britain had a bigger air force ratio when they won than when they lost, the differences are small.

d. Tank Ratio: "tank"

tank= T_A / T_D,

where

T_A = Number of tanks on the attacker side and

T_D = Number of tanks on the defender side.

Just like planes, tanks were used in WWI on a very small scale, but the real use of tanks happened in WWII.

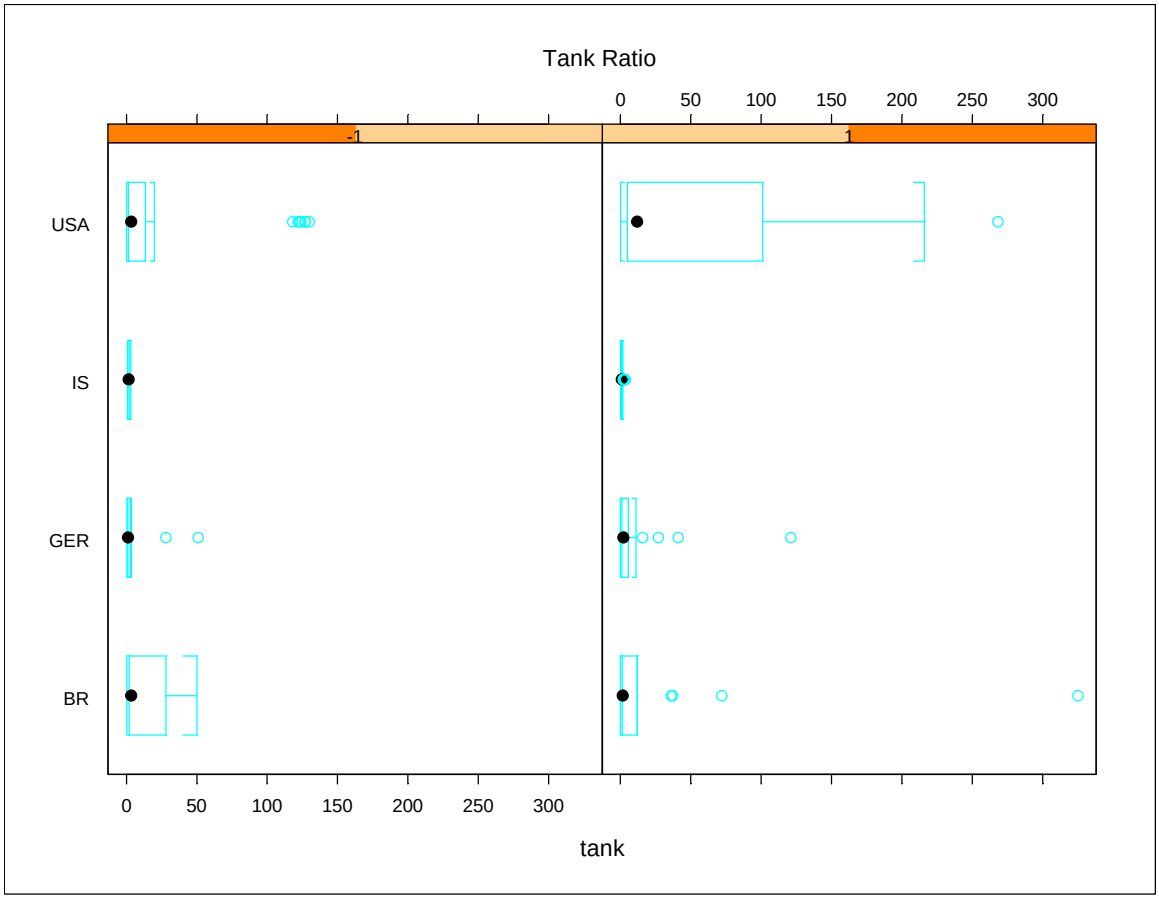


Figure 6. Tank Ratio, All Battles.

Figure 6 has battles of the entire data set. It is very similar to the plot drawn with the data from WWII [App B.D.]. This plot is also affected by the USA's dominance.

To compare countries other than the USA, the following boxplot is used.

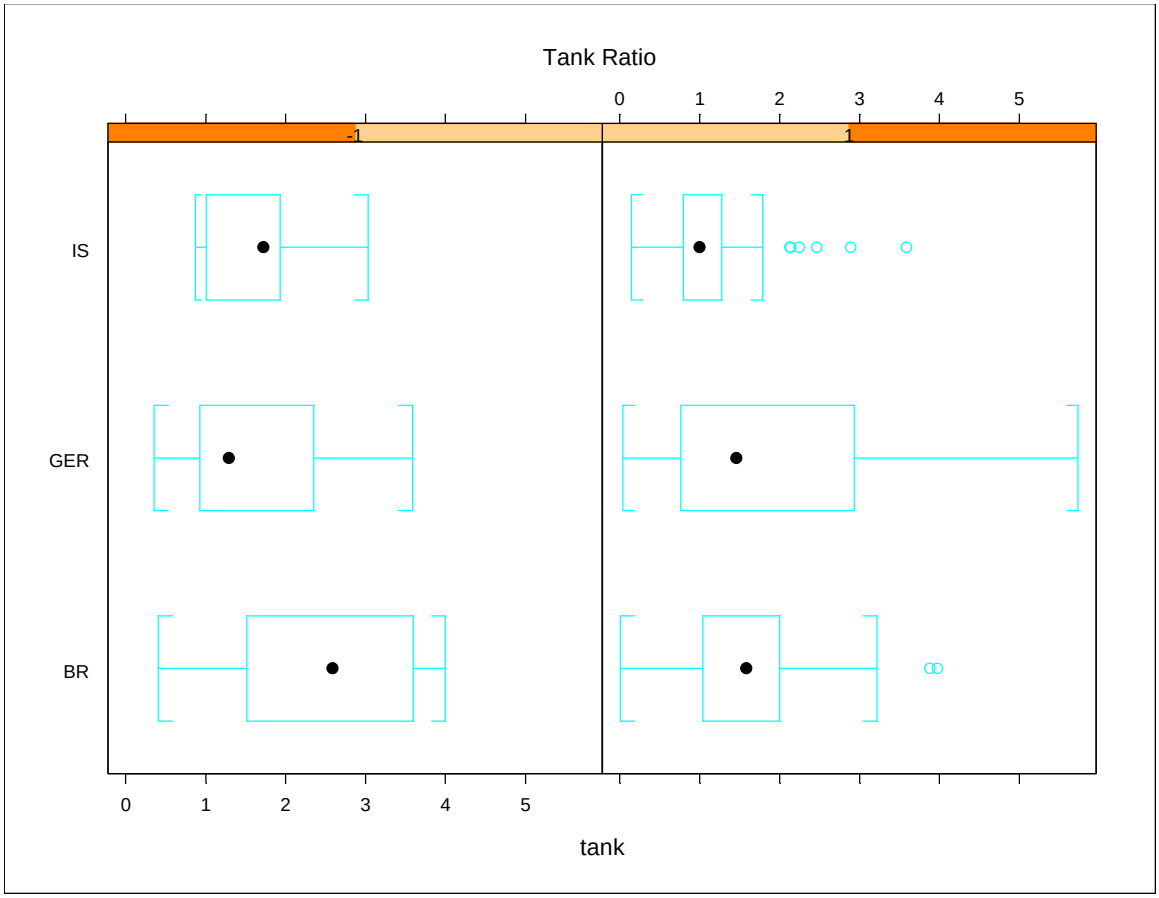


Figure 7. Tank Ratio, Israel, Germany and Britain

Figure 7 includes data with a tank ratio of 20 or less. With this truncation, 12 data points out of 501 are lost from Figure 6. This truncation is necessary to be able to compare these three countries.

According to Figure 7, Germany had a higher tank ratio than the other two countries. Britain and Israel's tank ratios, Britain's especially, were higher in the battles they lost than the battles they won. This variable seems to have different patterns within every individual country.

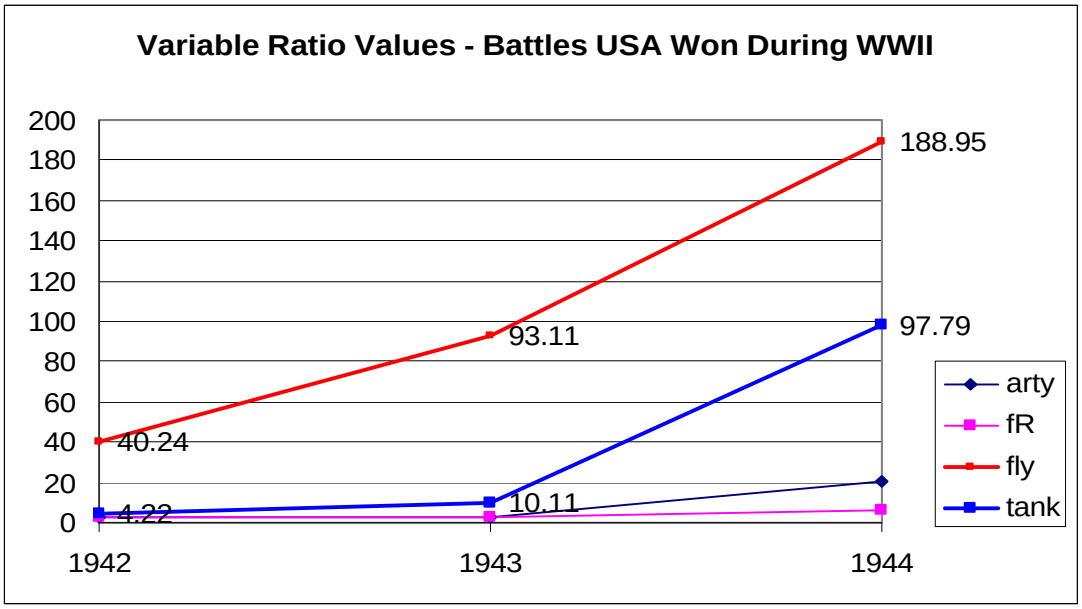


Figure 8. Ratios of the Objective Variables of Battles in WW2. The USA is the Attacker and Winner.

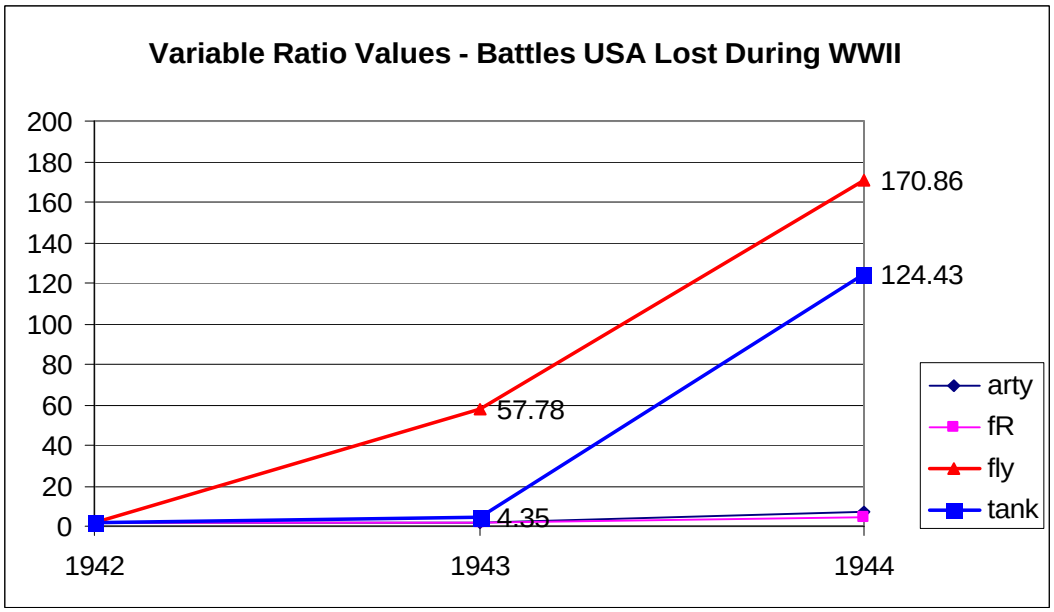


Figure 9. Ratios of the Objective Variables of Battles in WW2. The USA is the Attacker and Loser.

As Figures 7 and 8 suggest, towards the end of WWII, the USA began to have a very big advantage over its opponents. This advantage became very overwhelming with "tank" and "fly" ratios. This big advantage over the opponents is the main reason for the variability pattern in the charts analyzed above.

e. Cavalry Ratio: "cav"

cav = C_A / C_D

where

C_A = Number of cavalries on the attacker side and

C_D = Number of cavalries on the defender side.

Cavalry Ratio is present in the data set from 1600 to 1905.

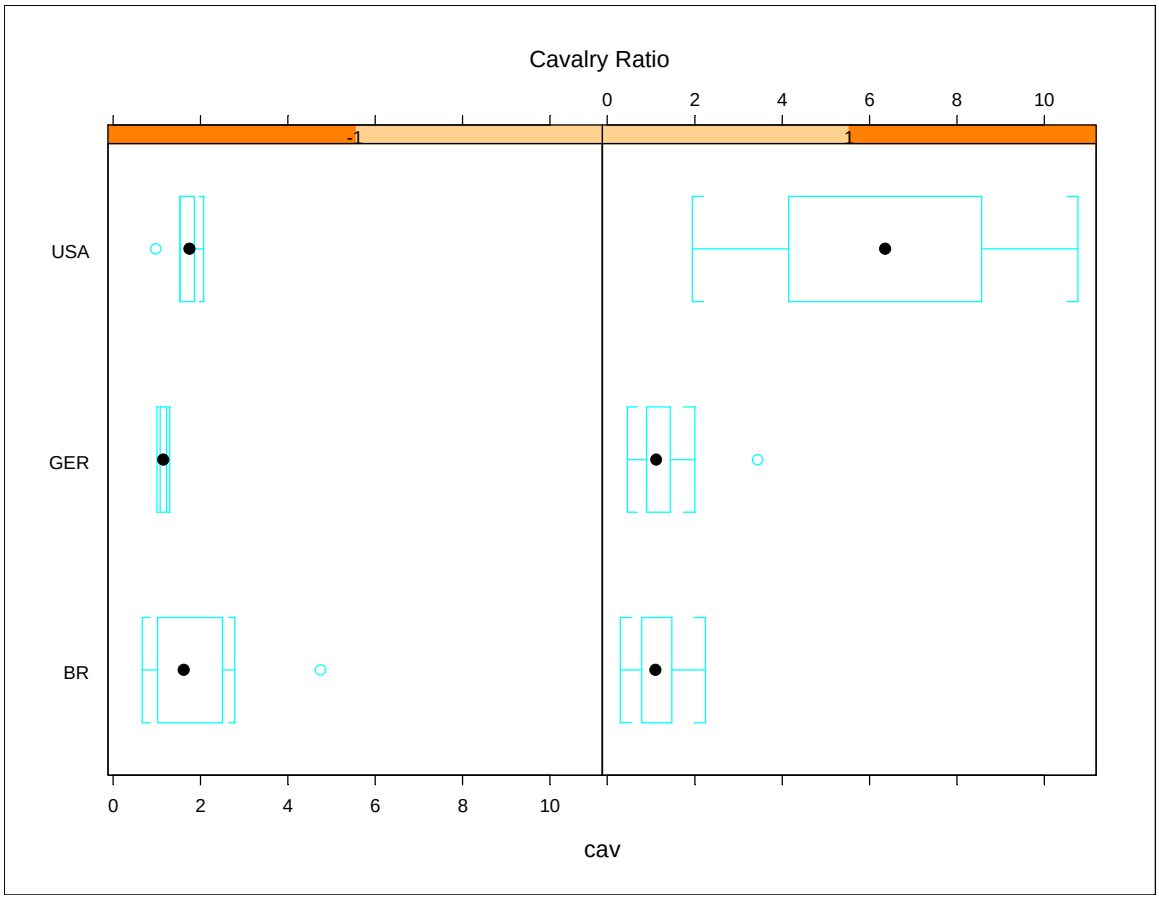


Figure 10. Cavalry Ratio.

While Britain and Germany had similar cavalry ratios when they won or lost, the USA had a much bigger ratio when it won than when it lost. The USA's cavalry ratio also has a high variability in the battles where the USA won. Again, the USA has a higher ratio than the other countries.

The cavalry ratio is the last objective variable to be analyzed. In the next section, the relative variables will be analyzed. A general discussion on all of the variables analyzed, both objective and relative, can be found at the end of this chapter.

4. Relative Variables

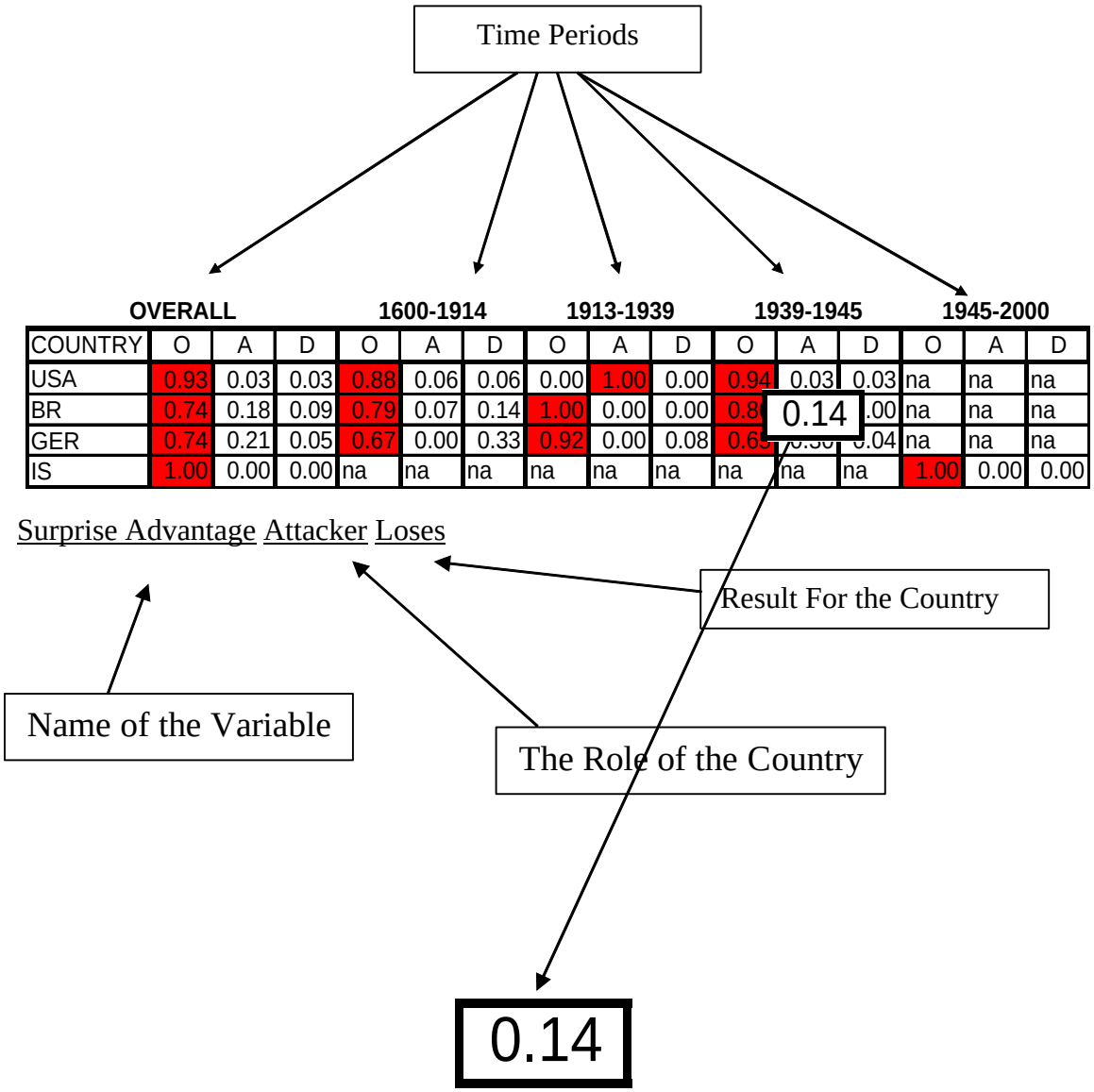
Relative variables are represented as categorical variables. As discussed in the previous section, relative variables are generally avoided by analysts in their models. Although no question exists concerning the importance of these variables, the fact that their values depend largely on personal judgment makes them less reliable.

Another reason for not using them in our classification trees is that the data does not have sufficient information on them. This is discussed at the end of this chapter in the "Discussion on the Relative Variables" section. However, a preliminary analysis is done to see the relationship between nationality and these variables. At this point, it is worth remembering our goal: to replace the relative variables with one variable: nationality.

In the following tables, the letter "A" denotes the battles in which the attacking side had an advantage, "D" denotes an advantage for the defending side and "O" means there was no advantage on either side. For example, if the "SURPA" is "A" for a particular battle, it means that, in that battle, the attacker had the "Relative Surprise" advantage, "D" says the defender had the advantage, and "O" says neither had the advantage. To familiarize the reader, an example table is explained below:

The "COUNTRY" column shows the name of the country. The "OVERALL" section of the time periods represents the whole dataset.

The cells having a number higher than 50 percent are highlighted, which makes it easier to see higher figures and patterns in the data. The cells where a corresponding figure is not available in the data set, for example, "ISRAEL" does not have any battles prior to 1946, and are simply marked with "na".



Now, we will explain how to read this table, using an example cell. The cell above says that, in WWII, among those battles in which Britain was the attacker and loser, the attacker (Britain) had the "Surprise Advantage" 14 percent of the time.

The relative variables considered important are analyzed below. First, we highlight the ones that appeared to be more important than the others. These two variables were also found important by Coban [Ref. 4].

a. Relative Surprise: "SURPA"

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.80	0.20	0.00	0.80	0.20	0.00	0.61	0.39	0.00	0.90	0.10	0.00	na	na	na
BR	0.66	0.34	0.00	0.75	0.25	0.00	0.45	0.55	0.00	0.71	0.29	0.00	na	na	na
GER	0.57	0.43	0.00	0.79	0.21	0.00	0.43	0.57	0.00	0.45	0.55	0.00	na	na	na
IS	0.55	0.45	0.00	na	na	na	na	na	na	na	na	na	0.55	0.45	0.00

Table 6. Surprise Advantage Attacker Wins.

For most of the battles, regardless of nation or period, there was no advantage on either side. Significant ones are highlighted. It is worth noting that in all the battles the attacker won, the defender *never* had a surprise advantage.

b. Relative Initiative Advantage: "INITA"

This is one of the more important variables. It can be said that, among all the relative variables, this is the only one with consistently significant values. One interesting point is that the attacker had an initiative advantage in more than 75 percent of the battles it won in all of the subsets except for one. Germany had an advantage in 64 percent of the battles during WWI. The defender never had an initiative advantage, zero percent of the time, when the attacker won.

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.14	0.86	0.00	0.05	0.95	0.00	0.03	0.97	0.00	0.22	0.78	0.00	na	na	na
BR	0.20	0.80	0.00	0.25	0.75	0.00	0.18	0.82	0.00	0.18	0.82	0.00	na	na	na
GER	0.15	0.85	0.00	0.16	0.84	0.00	0.36	0.64	0.00	0.00	1.00	0.00	na	na	na
IS	0.10	0.90	0.00	na	na	na	na	na	na	na	na	na	0.10	0.90	0.00

Table 7. Relative Initiative Advantage, Attacker Wins.

SURPA and INITA are the two of the most important variables in the data set. Other variables will be discussed with the help of Table 8. This table includes all of the relative variables used in the analysis. Time periods are not as detailed as in the

individual tables like Tables 7 and 8. Tables of individual variables are in Appendix A, and they will be referred to when necessary. Again, all of the figures of Table 8 are from the battles where the countries were attacking.

In Table 8, the values in the cells, as in previous tables, are the proportions. To make the tables easier to read, the cells are formatted as follows:

: If the cell contains a value greater than or equal to 0.8

: If the cell contains a value between 0.5 and 0.8

: If the cell contains a value less than 0.2

Table 9 contains the exact numbers for each cell, instead of ratios. Figures in each cell refer to the number of battles.

		WIN			LOSE		
VARIABLE	COUNTRY	O	A	D	O	A	D
SURPA	USA	0.80	0.20	0.00	0.93	0.03	0.03
	BR	0.66	0.34	0.00	0.74	0.18	0.09
	GER	0.57	0.43	0.00	0.74	0.21	0.05
	IS	0.55	0.45	0.00	1.00	0.00	0.00
CEA	USA	0.85	0.13	0.03	0.72	0.05	0.23
	BR	0.48	0.32	0.20	0.85	0.03	0.12
	GER	0.51	0.49	0.00	0.71	0.24	0.05
	IS	0.00	1.00	0.00	0.14	0.86	0.00
LEADA	USA	0.87	0.11	0.02	0.69	0.00	0.31
	BR	0.68	0.30	0.02	0.68	0.03	0.29
	GER	0.53	0.47	0.00	0.82	0.08	0.11
	IS	0.03	0.97	0.00	0.43	0.57	0.00
TRNGA	USA	0.64	0.12	0.25	0.69	0.03	0.28
	BR	0.57	0.23	0.20	0.76	0.12	0.12
	GER	0.64	0.34	0.02	0.71	0.26	0.03
	IS	0.00	1.00	0.00	0.43	0.57	0.00
MORALA	USA	0.55	0.45	0.00	0.77	0.23	0.00
	BR	0.89	0.11	0.00	0.97	0.03	0.00
	GER	0.87	0.13	0.00	0.92	0.03	0.05
	IS	0.66	0.34	0.00	0.57	0.29	0.14
LOGSA	USA	0.85	0.14	0.02	0.92	0.03	0.05
	BR	0.86	0.09	0.05	0.85	0.06	0.09
	GER	0.85	0.09	0.06	0.87	0.03	0.11
	IS	1.00	0.00	0.00	1.00	0.00	0.00
MOMNTA	USA	0.67	0.33	0.00	0.89	0.08	0.03
	BR	0.80	0.20	0.00	0.85	0.15	0.00
	GER	0.64	0.36	0.00	0.76	0.24	0.00
	IS	0.55	0.45	0.00	0.57	0.29	0.14
INTELA	USA	0.95	0.04	0.02	0.82	0.02	0.16
	BR	0.82	0.14	0.05	0.85	0.00	0.15
	GER	0.60	0.32	0.08	0.79	0.05	0.16
	IS	0.93	0.07	0.00	0.86	0.00	0.14
TECHA	USA	0.80	0.20	0.00	0.93	0.07	0.00
	BR	0.66	0.34	0.00	0.97	0.03	0.00
	GER	0.57	0.43	0.00	0.95	0.05	0.00
	IS	0.55	0.45	0.00	1.00	0.00	0.00
INITA	USA	0.14	0.86	0.00	0.39	0.48	0.13
	BR	0.20	0.80	0.00	0.53	0.38	0.09
	GER	0.15	0.85	0.00	0.39	0.53	0.08
	IS	0.10	0.90	0.00	0.29	0.71	0.00
AEROA	USA	0.63	0.09	0.29	0.00	0.97	0.03
	BR	0.57	0.00	0.43	0.00	1.00	0.00
	GER	0.65	0.30	0.04	0.13	0.04	0.83
	IS	0.14	0.86	0.00	0.14	0.86	0.00

Table 8. Ratio of All Relative Variables.

		WIN			LOSE		
VARIABLE	COUNTRY	O	A	D	O	A	D
SURPA	USA	88	22	0	57	2	2
	BR	29	15	0	25	6	3
	GER	30	23	0	28	8	2
	IS	16	13	0	7	0	0
CEA	USA	93	14	3	44	3	14
	BR	21	14	9	29	1	4
	GER	27	26	0	27	9	2
	IS	0	29	0	1	6	0
LEADA	USA	96	12	2	42	0	19
	BR	30	13	1	23	1	10
	GER	28	25	0	31	3	4
	IS	1	28	0	3	4	0
TRNGA	USA	70	13	27	42	0	19
	BR	25	10	9	23	1	10
	GER	34	18	1	31	3	4
	IS	0	29	0	3	4	0
MORALA	USA	60	50	0	47	14	0
	BR	39	5	0	33	1	0
	GER	46	7	0	35	1	2
	IS	19	10	0	4	2	1
LOGSA	USA	93	15	2	42	0	19
	BR	38	4	2	23	1	10
	GER	45	5	3	31	3	4
	IS	29	0	0	3	4	0
MOMNTA	USA	74	36	0	54	5	2
	BR	35	9	0	29	5	0
	GER	34	19	0	29	9	0
	IS	16	13	0	4	2	1
INTELA	USA	104	4	2	50	1	10
	BR	36	6	2	29	0	5
	GER	32	17	4	30	2	6
	IS	27	2	0	6	0	1
TECHA	USA	88	22	0	57	4	0
	BR	29	15	0	33	1	0
	GER	30	23	0	36	2	0
	IS	16	13	0	7	0	0
INITA	USA	15	95	0	24	29	8
	BR	9	35	0	18	13	3
	GER	8	45	0	15	20	3
	IS	3	26	0	2	5	0
AEROA	USA	22	3	10	0	34	1
	BR	4	0	3	0	7	0
	GER	15	7	1	3	1	19
	IS	1	6	0	1	6	0

Table 9. All Relative Variables with the Number of Battles.

Tables 8 and 9 summarize all relative variables for our four countries while attacking. There is no time segmentation. All the sections of the table reveals figures from the entire dataset. While Table 8 contains ratios, Table 9 shows the exact number of battles in each cell.

The rest of the relative variables, those with less significance, are listed below. The detailed time period tables are provided in Appendix A.

(1) Relative Combat Effectiveness: "CEA". Until WWII, when the attacker won, the defender never had a combat effectiveness advantage. In WWII, in Britain's battles, the defender had this advantage in 53 percent of the battles and Britain still won. Israel had this advantage even in the battles it lost (86 percent). [App A.B.]

(2) Relative Leadership Advantage: "LEADA". Until WWI, in more than half of the battles, the side with this advantage won. In WWI and WWII, neither side had a significant advantage. In the battles Israel fought, the defender never had a leadership advantage. [App A.D.]

(3) Relative Moral Advantage: "MORALA".. There is no significant moral advantage on either attacker or defender side except the USA. In WWI, the USA had relative moral advantage in all battles. [App. A.F.]

(4) Relative Logistics Advantage: "LOGSA" None of the nations had a significant logistics advantage in any of the battles [App A.G.]

(5) Relative Momentum Advantage: "MOMNTA". The only significant advantage is on Germany's side in WWII. Germany had a momentum advantage in 65 percent of the battles in WWII where it attacked and won. Also, in the entire data set, the defender never had the momentum advantage. The exception is US battles in WWII. In six percent of the battles, the USA lost when attacking and the defender had a momentum advantage. [App A.H.]

(6) Relative Intelligence Advantage: "INTELA". There is no significant advantage on either side. The only exception is Germany in both World Wars. It is not very significant, but when Germany attacked and won, it had an advantage in 36 percent of the battles in WWI and 40 percent in WWII. [App A.I.]

(7) Relative Air Superiority: "AEROA". This variable determines the quality of the air force. The USA and Britain had this advantage in a large portion of the battles they lost, but not so large in the battles they won. Israel had this advantage in 86 percent of its battles

C. GENERAL DISCUSSION ON RELATIVE VARIABLES

In this section, we look at the relative variables together.

Looking at Table 8, three trends can be easily seen:

- The defender hardly ever had an advantage over the attacker. The countries we analyze are the attackers. Only 7 out of 80 cells belonging to the defender have values of more than 20 percent, the highest being 31 percent. Thus, can we say that this data suggests these countries always fought with the countries possessing inferior qualities? Not really, because they also fought with each other. However, interestingly enough, according to the data, for all variables with the exception of "Initiative Advantage", there is often no advantage on either side. In more than 50 percent of the battles, neither side had the advantage, except for those of Israel.
- Israel has different values than other countries analyzed. For the variables "CEA", "LEADA" and "TRNGA", Israel had an obvious advantage over the defender. However, interestingly enough, there is no significant difference between Israel's degree of advantage when they won or lost. For example, they had a Leadership advantage in 57 percent of the battles both when they won and they lost. And, they had Combat Effectiveness advantage in all of the battles, 100 percent, when they won and 86 percent of the battles they lost.
- "Initiative Advantage" is the only variable where the attacker consistently had an advantage over the defender. This results from the fact that the attack is often done to seize the initiative. "Offensive operations are the means by which a military force seizes and holds the initiative while maintaining freedom of action and achieving decisive results. This is fundamentally true across all levels of war." [Ref. 11]. In other words, the attacking side has the initiative advantage almost "by definition".
- If the two exceptions discussed earlier, Israel from the countries and "Initiative Advantage" from the variables, are put aside, in more then 50 percent of the battles, there is no advantage on the either side. As an attacker, Britain had a Combat Effectiveness Advantage in 48 percent of the battles, which is the only exception. This also supports our earlier claims. To decide the values of relative variables is so difficult that even the historians could not find an advantage on either side in more than 50 percent of the battles.

D. SUMMARY

In this chapter, the variables that are considered to be important were analyzed with respect to the countries. Other than the results discussed in the previous sections, some other important considerations are given below:

- It is important to note that, although this data set is the best data set on historical land combat, it is not at all perfect. It would be a serious mistake to accept this data as the ultimate truth because:
  1. The data was collected by military historians. Therefore, the battles listed in the data set are decided upon their comfort level. They are not all of the battles fought, nor are they necessarily the most important ones or a random sample. It may be the case that they did not have sufficient data on many very important battles and therefore ignored them.
  2. The countries we focus on, the USA, Britain, Germany and Israel, are usually considered to be successful on the battlefield. We are forced to do this because these are the ones for which sufficient data exists. They all have similar kinds of properties: Extensive use of technology, a large economy behind the war machine and extensive experience. Some may argue that Israel does not fall into that category, but compared to their enemies, the difference is obvious. As a result, it is hard to find the advantages caused by nationality.
- All of the battles that were in the data set have only one nation as the attacker and the defender. It is a fact that this is not the case in many battles. There were and still are alliances. This appears to be another limitation of the data set.
- An interesting point discussed before is the fact that the USA had a huge power on the field towards the end of WWII. It really is difficult to analyze the nationality factor of the USA. It can be said that accumulating a big power on the battlefield and outnumbering the enemy is a main characteristic of the USA, but this could only be decided upon for certain with the support of military historians.
- When the objective variables were analyzed, all countries had different characteristics. Had they had similar properties, it would have been easier to determine the effect of nationality. However, in our case, one may not be able to decide whether it is the objective variables or the nationality factors that affects the outcome.
- Israel’s smaller values than the others, smaller force ratio, artillery ratio etc., suggest that the way in which wars are fought has changed. Even fewer weapons can and do provide more lethality.

THIS PAGE INTENTIONALLY LEFT BLANK

III. CLASSIFICATION TREES

A. INTRODUCTION

In this chapter, classification tree models will be analyzed. First, the reader will be informed about tree-based modeling, a relatively new analysis method. Second, we will introduce the tree models built using the CDB90G data set. A discussion on the models built, and further study suggestions, will conclude this chapter.

Tree-based modeling is an exploratory technique for uncovering structure in data. Specifically, the technique is useful for classification and regression problems when one has a set of classification or predictor variables (x) and a single response variable (y). Tree-based models are relatively new, but are gaining widespread popularity as a means of devising prediction rules for rapid and repeated evaluation, as a screening method for variables, as a diagnostic technique to assess the adequacy of linear models, and simply summarize large multivariate datasets. [Ref. 7] We will use them for the latter purpose.

Trees simply show the structure of the data. Trees do not need distributional assumptions, and as such, transformations are not needed. Any interactions between variables are automatically included in the tree structure. Furthermore, they are robust to outlying data.

Trees are arranged hierarchically. Until a terminal node is reached, the data flowing down the tree encounters one decision at a time.

One of the advantages of tree-based models is that they are easy to read. There are oval (non terminal or split) and rectangular (terminal) nodes. Each node contains the predicted outcome and the distribution to the child nodes. The split criterion is shown on each branch.

For example, Figure 11 shows a tree model built on the entire data set. The *root* node says that there are 657(260+397) data points in the data set. 260 of them are the ones where the attacker did not win and 397 times the attacker won. The first split is determined by which country is attacking. If the defender is Britain, the Confederate States, Israel, or the USA, we go to the left node (a *terminal* node) and if the defender is one of the other nations, Austria, Egypt or the other ones mentioned in the right branch

go to the right node (which is a *split* node) on the right. According to the *terminal* node on the left, the tree model predicts that the attacker does not win, i.e. it could be a loss or a draw. At the terminal node on the left, there are 156 observations, 87 of which are "-1" for attacker losses, and 51 are "1" for attacker wins. If we go to the right branch, we reach a split node. In that split node, there are 173 observations with a WINA value of "-1" and 346 observations with a WINA value of "1". At that node, the question is "What is the force ratio?" If the force ratio is less than 5.38194, then the attacker is predicted to win. This is the terminal node in the middle. Again, 171 refers to the number of "-1"s in that node, and 318 refers to "1"s. If the force ratio is greater than 5.3819, the right branch is chosen, which also suggests a win for the attacker. However, as the reader can recognize, although both of the terminal nodes, the middle one and the one on the right, suggest a win for the attacker, the misclassification rates are different. Now, we will discuss the algorithm behind the tree models and how the splits are decided and the tree built.

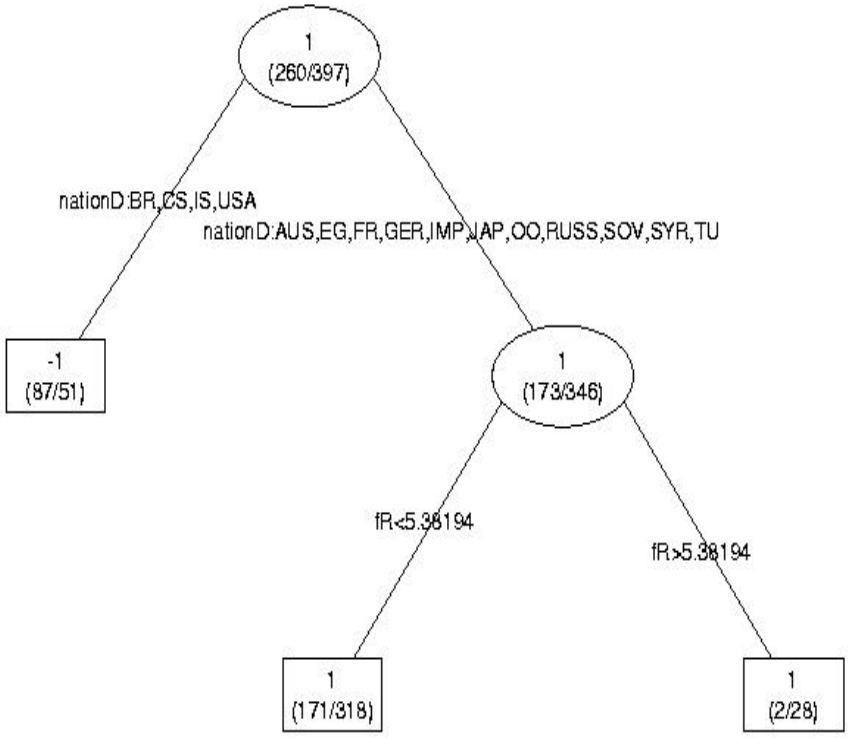


Figure 11. Tree Model of the Entire Data Set.

The tree models are fit by binary recursive partitioning, by which the data set is successively split into increasingly homogenous subsets. [Ref. 7] The usual set-up for regression, or classification if the response variable is categorical, trees is as follows. The *n* responses, in our models the variable WINA, $y_1,\ldots, y_n$, and the predictors x_i are collected for each y_i. Starting with all *y*'s in one node, the *impurity* of that node is measured. *Impurity* can be one of several different measures, deviance or Residual Sum of Squares (RSS), Mean Sum of Squares (MSE) etc. Both S+ and the `rpart` algorithm measure *impurity* by deviance. For more information on deviance, see Devore pp. 502-503 [Ref. 8]. The objective is to divide the observations into sub-nodes of high purity, i.e., have as many similar *y*'s as possible. So, at a node (a "split"), if the data is categorical, the data is

divided into two subsets one including some of the categories, the other with the rest, e.g., we might separate a group of nations from the others. If the data is continuous, every possible split of the form X<a is considered. The criterion by which the split is decided is called **the split criterion**. Then, the *impurity* (RSS) is computed for each of the two groups. The split decreasing the impurity most is chosen. [Ref. 12]. This process of splitting can continue down to every single observation, which would be over-fitting. In order to avoid this, the tree construction continues until the number of observations in each node is small, by default n_i<20 for `rpart,` or the leaf is sufficiently homogenous, i.e., with small *impurity*.

There are several tree methods available. Since there are many missing values in the data set (Table 10), we prefer to use the `rpart` [Ref. 10] method because of the way it handles the missing values.

PERIOD	fR	arty	fly	cav	tank	Total Number of Battles
Before WWI	0	95	251	107	251	251
WWI	0	46	126	130	101	133
WWII	1	7	81	193	5	193
After WWII	0	2	22	79	4	80
TOTAL	1	150	480	509	361	

Table 10. Number of Missing Values.

This table gives the number of missing values for the objective variables used in building the tree models. It is important to note that tanks and airplanes were not present before WWII and they were used on a very small scale during WWI. Also, cavalry was used mainly before WWI. Therefore, some of the big numbers are basically historical facts. However, even taking this into consideration, a large missing value problem exists forcing us to use `rpart`.

In `rpart`, when missing values are encountered in considering a split, they are ignored and the probabilities and *impurity* measures are calculated from the non missing values of that variable. Surrogate splits are then used to allocate the missing cases to the daughter node. Therneau [Ref. 10] contains some more detail about the usage of surrogate splits in `rpart`. The next two paragraphs briefly explain the use of surrogate splits.

Once a splitting variable and a split point for it have been decided, what is to be done with observations missing that variable? One approach is to estimate the missing datum using the other independent variables. `rpart` uses a variation of this to define surrogate variables.

As an example, assume that "Force Ratio <2" has been chosen as the split criterion and there are data points missing information on the force ratio. The surrogate variables to be used for the data points which are missing the value for the force ratio, are then found by re-applying the partitioning algorithm (without recursion). The two categories "Force Ratio <2", "Force Ratio >2" are predicted using the other independent variables. For each predictor, an optimal split point and a misclassification error are computed. The surrogates are than ranked. Any observation that is missing the split variable is then classified using the first surrogate variable, or if that is missed, the second surrogate is used, and so forth. If an observation is missing all surrogates, the blind rule of "go with the majority" is used. Other strategies for these "missing everything" observations can be argued, but there should be few or no observations of this type. [Ref. 10]

Another issue with tree models is pruning. After building the model, it is usually the case that it is over-fitted. [Ref. 12] The trees are built in order to minimize the *impurity*. In doing so, the trees grow too big. In other words, the models are too good. This, of course, decreases the model's ability to predict. Pruning is used to solve this problem.

The question we ask then is whether we want a predictive model or a descriptive (explanatory) one. As mentioned above, trees can be used for both. If it is a predictive model, pruning to the optimum size using cross-validation is vital. However, for descriptive models, in other words, models to explore the data, pruning is not that great of a concern. The tree explaining the data best is used.

One of the problems faced in this situation is lack of sufficient data to build predictive models. When trees we built are pruned to the optimum size, they become too small to be used in predictions. As a result, we will use tree models to describe the data and explore the nationality factors in the data set. For this reason, pruning will not be our concern. All the models built are descriptive models.

B. TREE MODELS

In this section, the tree models that were built are examined. Trees were built by using the battles in which the countries that we are analyzing, the USA, Britain, Germany, and Israel, appeared either as an attacker or a defender, and only the objective variables were used as predictive variables. Since not all of the variables are present in each period, only the appropriate ones, whose names are given in the sections where the trees described are used to build the models.

1. Model 1: The Battles Prior to World War I

This is the model for the battles before 1910, see Figure 12. The model shows that nationality was the most important factor affecting the outcome of the battle. In other words, if only one variable were allowed, we would choose: "What is the nationality of the attacker?" Three of the countries in which we are interested appeared at the first split. According to the model, the USA, Germany and Britain tended to win the battles in which they were defending. This was correct in 47 of the 69 battles in which they were defending.

2. Model 2: The Battles of World War I

This is the model for the battles of WW I, see Figure 13. According to the model, the most important factor was force ratio. The second most important was nationality. However, it is important to note that the split criterion for the force ratio is 4.05. There are 15 battles where the attacker had a ratio at least 4.05 and the attacker won them all. Out of these 15, 10 were from the USA, 3 from Germany and 2 from Britain. This, again, leads us to the same question asked previously: Is it the nationality or the objective factors that have the real effect?

3. Model 3: The Battles of World War II

During WW II, the most important variable was artillery ratio. The second most important was, as in WWI, nationality. The USA, Germany and Britain again appear in the second split. They won the battles in which they were defending against an attacker who did not have sufficient artillery support.

4. Model 4: The Battles that Israel Fought

This particular model follows our historical segmentation. This model contains the battles fought after WWII, but instead of including all the battles, we focused on Israel, and tried to ascertain if nationality factors are important.

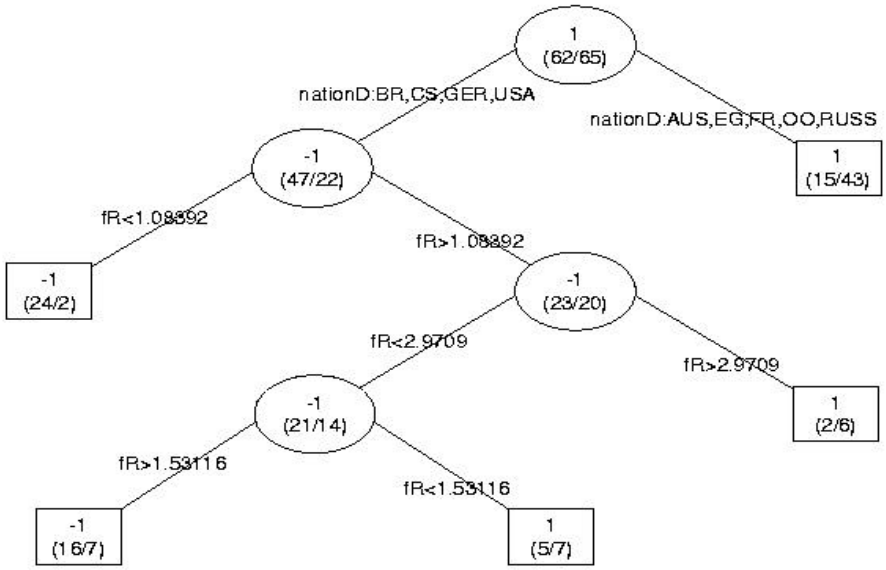


Figure 12. Model 1 Battles Before World War I.

Model 1 explains 76 percent of the battles. That is, the terminal nodes correctly classify the outcome 76 percent of the time. The most important factor is nationality. The USA, Britain and Germany appear in the first split as defenders explicitly and as attackers implicitly. Implicitly, because, these are the battles of those three countries. When other countries are defending, they are the attackers. If we were to predict an outcome of a hypothetical battle in this period, we could predict the result only by looking at the nationality of the countries and we would be correct 71 percent of the time. If the USA, Britain or Germany is either defending or attacking, they win. With the exception of a few draws, a value of "-1" refers either to a draw or a loss for the attacker. After nationality, the single most important variable is force ratio. Other variables present at this period, artillery ratio and cavalry ratio, did not appear in our tree. The first split including force ratio is 1.08, which is interestingly small. This reminds us of the findings of Yigit [Ref. 3], and his work on the 3 to 1 force ratio rule-of-thumb. As can be seen from the force ratio boxplots [App. 1.1], before WWI, the force ratios are small and there is no significant difference between the force ratios of the countries analyzed. This helps us to understand two things: (1) As we claimed before (Chapter II Conclusions) since the countries have similar properties, it is easier to decide whether nationality has an important effect. Also, in this case, the tree decided that nationality is the primary factor. (2) The split criteria related to force ratio are small because all countries had similar force ratios.

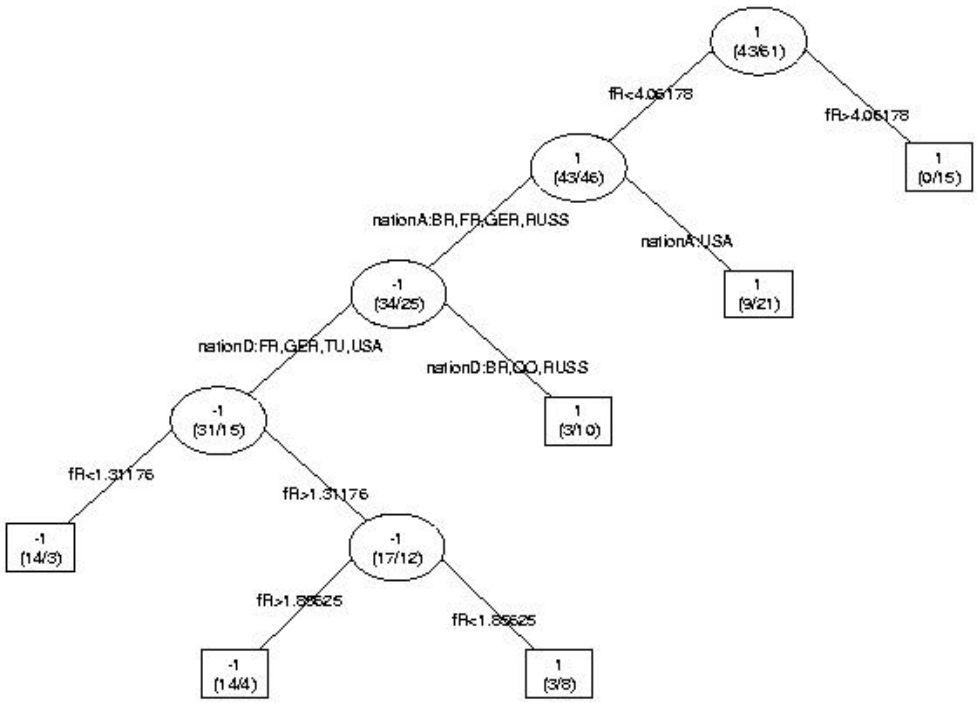


Figure 13. Model 2 Battles of World War I.

This model explains 79 percent of the battles. The most important factor is force ratio. The second most important factor is nationality. In this period, force ratio began to be much more important on the battlefield. Also, now, unlike the battles prior to WWI, the threshold is much higher. As the reader may recall, the first threshold for force ratio in the battles prior to WW I was 1.08, see Figure 12, as opposed to 4.06 in WWI. This is mostly because of the USA's high force ratios. 10 out of 15 observations in the terminal node on the right is from the battles of the USA. Also, this is not surprising, considering the fierce defenses of that era. It takes mere power, i.e., force ratio, to defeat the defender. Another point, though, all 15 battles that have the large force ratio have the tree countries as the attacker. Again, how can one decide whether it is the force ratio or nationality that affects the outcome? Second and third splits are nationalities. If the USA is attacking, they won even with a force ratio less than 4.06. Britain and Germany are less successful at attacking, but the Germans were better defenders than the Britons. The USA is good at both defending and attacking.

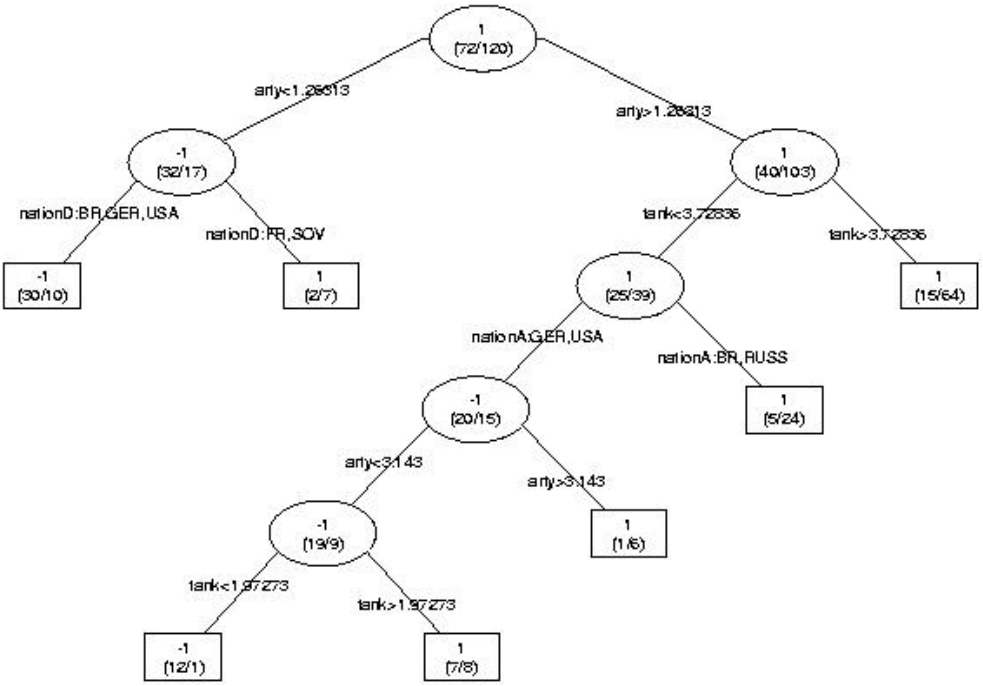


Figure 14. Model 3, Battles of World War II.

This model explains 79 percent of the battles. The most important variable is artillery ratio. Technology and advanced weapons began to play a more important role on the battlefield. In the battles where the attacker did not have an artillery ratio advantage, the second most important factor is nationality. In those battles, the USA, Britain and Germany won as defending armies. In the battles where the attacker has an artillery advantage, the second most important variable is tank ratio. An advantage of 3.7, along with an artillery advantage of 1.3, almost guaranteed the attacker's victory, 64 out of 79 battles. When the tank ratio is smaller, Britain won the battles where they attacked, while Germany and the USA lost as attackers, if they do not have a tank ratio of 1.9 or an artillery ratio of 3.15 or more. To summarize, all three countries are good defenders; Britain is a better attacker when they have less power than Germany and the USA. However, the importance of weapons appears much higher than in the previous time periods.

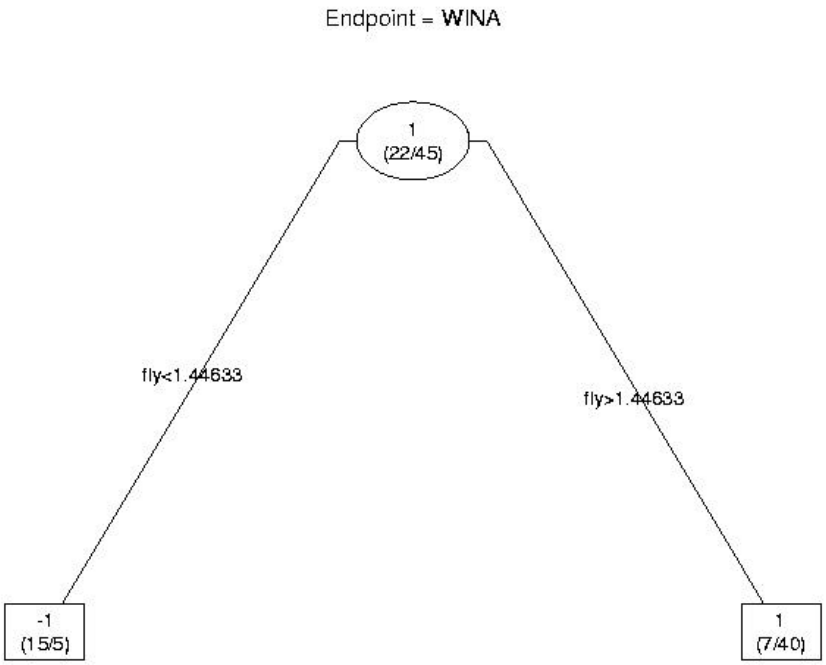


Figure 15. Model 4, Battles that Israel Fought.

This simple model explains 81 percent of the battles. The importance of advanced weapons is still increasing. The only important variable is air force ratio. However, we know from the data set that Israel won 82 percent of the battles they attacked. Again, as we claimed previously, see Chapter II, Conclusions, the countries we are analyzing are those already using the more important factors, the decisive variables such as artillery and tank ratio in WW II or air force ratio after WW II. Thus, it is difficult to decide where nationality has a role, or what is a nationality factor. These issues will be addressed at the end of this chapter.

After the models are fit, the question is how good are the models, or how important are the nationality factors? For a predictive model, there are a couple ways to ensure quality. One is cross-validation, which is used with `rpart`'s "`prune.rpart()`" command. This was tried and proved ineffective because of a lack of information. The shortage of data, as discussed above, was the main reason for building explanatory models rather than predictive ones. Another way to build better trees is by dividing the data in random subsets, and then building the model using one of the training set subsets. After building the tree, it is evaluated with the rest of the data.

Doing this with different subset variations until a good tree is built is another approach to build good models. However, we have the same problem with cross-validation: insufficient information.

Our models, as mentioned previously, are explanatory as opposed to predictive models. Thus, there is another measure on which we can assess our models: the misclassification rate.

The misclassification rate is the measurement of what percentage of the data can actually be explained with the model. We will use it as *our* measurement to assess the models. Our models with the nationalities were presented in the previous section. To evaluate the importance of nationality, models without the nationality factors were also built. They will not be presented, but instead, misclassification rates with and without nationality will be compared. The next table contains those values.

	WITH NATIONALITY	WITHOUT NATIONALITY
BEFORE WW1	0.244	0.315
DURING WWI	0.212	0.250
DURING WWII	0.216	0.219

Table 11. Misclassification Rates of the Trees with and without Nationality.

As can be seen from the table, the effect of nationality, the change in the misclassification rate with nationality, was largest before WWI. That is not a surprise, since nationality was the primary split in that period. An approximate 7 percent improvement in the misclassification rate occurred when nationality is used. During WWI, the improvement was 3.8 percent. Beginning with WWII, the nationality variable began to be a rather unimportant factor.

C. SUMMARY

In this chapter, we analyzed the data set using the classification trees. Some of the important conclusions reached follow.

- Nationality is the most important variable prior to WW I.
- There is an obvious trend in history related to predicting the outcome of the battle, that is, as time passes. Technology and advanced weapons play a more important role in deciding the outcome of the battle. Force ratio was the decisive factor up to WW II, but artillery and tank ratios were in WW II and the air force ratio after that. However, as demonstrated, the countries we analyzed, the USA, Germany, Britain and Israel usually use

those weapons more effectively than the others. This, their consistent ability to use the most effective weapon systems, is their characteristic. Thus, even if the exact figures about their force structure are not available, it would not be wrong to predict that they have enough to win the battle they are fighting.

- The USA will almost certainly have an overwhelming force on the battleground to ensure they win.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. CONCLUSION

The analyses produced some interesting results. As we mentioned in the introduction chapter, our purpose was to find the importance of nationality on battle outcomes. For the analyses, we did the following:

- The analyses focused on four different countries: The USA, Germany, Israel and Britain.
- Since the nature of warfare evolves, the data set is divided into four periods: battles before WWI, WWI, WWII and the battles after WWII.

By combining our findings from summary statistics and the tree models, we conclude the following:

- Relative variables are avoided internationally and it is not a good idea to use them in the models. The reason for avoiding them is that they are subjective and hard to determine before a battle. In addition, we also found that the data set does not contain much information on the values of these variables. That is, according to the data, in the majority of the battles neither side has an advantage. In other words, even if one decides to use the relative variables in a model, it will be difficult to find discriminatory information in this data set.
- The tree models show that nationality was the most important factor in the battles before WWI. This is in line with the findings of Coban [Ref. 4], who found that in the battles before WWI, the relative variables are more important than objective variables. Here, we are using nationality as a surrogate for the relative factors. In this thesis, one of the questions we asked was whether we can replace all relative variables with just nationality. Also, as the results demonstrate, we can replace the relative variables with nationality alone, when relative variables are important. Coban's model for the battles before WWI has a misclassification rate of 21 percent, and ours has 24 percent. Although the analysis methods have minor differences in that his is a predictive model whereas ours is a explanatory one, the comparison provides a very good indication of the soundness of our model.
- The importance of weapons and technology has been increasing since the beginning of the 20^{th} century. Also, the countries examined made consistent use of weapons and technologies, which affect the outcome of the battle. Therefore, even though we cannot determine the exact importance of nationality by examining the results of our tree models, we can conclude that when combining them with other analyses, the four countries, the USA, Germany, Britain and Israel are expected to have sufficient weapons on the battleground to win the battle. Considering the amount of the data existing on the battles of the USA, it is easier for us to

reach a conclusion about the USA's nationality factor. That is, the USA almost always has had an overwhelming military power and this is a national characteristic of the USA. Looking at recent combats, this conclusion seems to be solid, even more so today.

- Although we conclude that it is the objective variables that are more associated with the outcome of the battle, we cannot say that an advantage in these guarantees success. The analyses in the second chapter showed that in most of the battles, no statistically significant difference exists between the relative variable values in the battles won or lost. This leads us to one truth about the phenomena of warfare: in war, luck and some other factors that can never be predicted nor can even be named, have a very big influence.

A. FURTHER STUDY SUGGESTIONS

- We used only the variable nationA, the nationality of the attacking country, as our response variable in the analyses done in Chapter II, Summary Statistics. It will be helpful to see what the results are also using nationD, the nationality of the defending country.

- Although S-Plus is a very powerful software package, it does have some limitations. Several new algorithms related to classification trees are available in other software packages. For example, with the methods available in S-Plus, each split has only two branches, but, in Clementine, the user can decide the number of branches at each split. It will be interesting to see what the results are if splits are forced on each nation, in other words, have the tree grow in such a way that every branch from a split has a different nation.

- With specific countries, further analyses can be done in more detail by using other statistical analysis techniques. For example, with the amount of data available, it is possible to analyze the battles of the USA and find their specific characteristics. Then, combining the results from these analyses, tree models can produce more significant results. Using cluster analyses might also be a good choice to analyze the data set.

- We talked about the data set and this data set not being the ultimate truth (Chapter II, Section D, Summary). Beyond that, in the analyses, we considered all battles equal, which in our opinion, is a pitfall. Battles are different, with respect to their size, or the importance of their results. In the same data set, a different selection of battles among all the others can be made. Professional help from a historian might be useful to do this. For example, having more homogeneous subsets and discarding the battles with an unreasonable force structure such as the ones in WW II in which the allies had an incredible advantage over Germany, may help reach better conclusions.

APPENDIX A. TABLES OF RELATIVE VARIABLES

In this section, tables for relative variables analyzed in the summary statistics section are provided. There are four tables for each variable, all with respect to the time periods. The first two tables are for the battles where the countries attack and win, the last two are for the battles where they attack and lose. The first and third tables have the exact number of battles for all countries. The second and fourth tables are with the four countries we analyze, and have the proportion of the battles' data. The reader may refer to p.26 for further explanation on how to read the tables correctly.

A. "SURPA"

SURPRIZE ADVANTAGE ATTACKER WINS

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	6	4	0	3	4	0	3	0	0	0	0	0	0	0	0
BR	24	13	0	7	2	0	5	6	0	12	5	0	0	0	0
CS	3	2	0	3	2	0	0	0	0	0	0	0	0	0	0
ENG	5	2	0	5	2	0	0	0	0	0	0	0	0	0	0
FR	25	11	0	21	10	0	4	1	0	0	0	0	0	0	0
GER	22	20	0	7	1	0	6	8	0	9	11	0	0	0	0
IS	16	13	0	0	0	0	0	0	0	0	0	0	16	13	0
OO	36	26	0	32	15	0	4	5	0	0	1	0	0	5	0
PR	8	3	0	8	3	0	0	0	0	0	0	0	0	0	0
RUSS	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0
SOV	16	6	0	0	0	0	0	1	0	16	5	0	0	0	0
USA	88	22	0	16	4	0	19	12	0	53	6	0	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.80	0.20	0.00	0.80	0.20	0.00	0.61	0.39	0.00	0.90	0.10	0.00	na	na	na
BR	0.66	0.34	0.00	0.75	0.25	0.00	0.45	0.55	0.00	0.71	0.29	0.00	na	na	na
GER	0.57	0.43	0.00	0.79	0.21	0.00	0.43	0.57	0.00	0.45	0.55	0.00	na	na	na
IS	0.55	0.45	0.00	na	na	na	na	na	na	na	na	na	0.55	0.45	0.00

SURPRIZE ADVANTAGE ATTACKER LOSES

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	11	1	0	7	1	0	4	0	0	0	0	0	0	0	0
BR	23	5	3	9	0	2	13	0	0	6	1	0	0	0	0
CS	11	6	0	11	6	0	0	0	0	0	0	0	0	0	0
ENG	2	1	0	2	1	0	0	0	0	0	0	0	0	0	0
FR	19	1	3	13	1	2	7	0	0	0	0	0	0	0	0
GER	26	8	1	0	0	0	11	0	1	15	7	1	0	0	0
IS	7	0	0	0	0	0	0	0	0	0	0	0	7	0	0
OO	40	7	4	21	4	3	8	0	0	2	0	0	9	2	1
PR	2	0	1	2	0	1	0	0	0	0	0	0	0	0	0
RUSS	4	1	0	2	1	0	2	0	0	0	0	0	0	0	0
SOV	4	0	0	0	0	0	2	0	0	2	0	0	0	0	0
USA	57	2	2	15	1	1	0	9	0	33	1	1	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.93	0.03	0.03	0.88	0.06	0.06	0.00	1.00	0.00	0.94	0.03	0.03	na	na	na
BR	0.74	0.18	0.09	0.79	0.07	0.14	1.00	0.00	0.00	0.86	0.14	0.00	na	na	na
GER	0.74	0.21	0.05	0.67	0.00	0.33	0.92	0.00	0.08	0.65	0.30	0.04	na	na	na
IS	1.00	0.00	0.00	na	na	na	na	na	na	na	na	na	1.00	0.00	0.00

B. "CEA"

RELATIVE COMBAT EFFECTIVENESS ATTACKER WINS

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	8	2	0	5	2	0	3	0	0	0	0	0	0	0	0
BR	16	12	9	3	6	0	5	6	0	8	0	9	0	0	0
CS	5	0	0	5	0	0	0	0	0	0	0	0	0	0	0
ENG	5	2	0	5	2	0	0	0	0	0	0	0	0	0	0
FR	23	10	3	18	10	3	5	0	0	0	0	0	0	0	0
GER	21	21	0	6	2	0	5	9	0	10	10	0	0	0	0
IS	0	29	0	0	0	0	0	0	0	0	0	0	0	29	0
OO	38	18	6	33	13	1	5	4	0	0	1	0	0	0	5
PR	6	5	0	6	5	0	0	0	0	0	0	0	0	0	0
RUSS	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0
SOV	9	7	6	0	0	0	1	0	0	8	7	6	0	0	0
USA	93	14	3	17	3	0	31	0	0	45	11	3	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.85	0.13	0.03	0.85	0.15	0.00	1.00	0.00	0.00	0.76	0.19	0.05	na	na	na
BR	0.48	0.32	0.20	0.50	0.50	0.00	0.45	0.55	0.00	0.47	0.00	0.53	na	na	na
GER	0.51	0.49	0.00	0.63	0.37	0.00	0.36	0.64	0.00	0.50	0.50	0.00	na	na	na
IS	0.00	1.00	0.00	na	na	na	na	na	na	na	na	na	0.00	1.00	0.00

RELATIVE COMBAT EFFECTIVENESS ATTACKER LOSES

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	9	0	3	5	0	3	4	0	0	0	0	0	0	0	0
BR	27	1	3	11	0	0	12	1	0	4	0	3	0	0	0
CS	17	0	0	17	0	0	0	0	0	0	0	0	0	0	0
ENG	2	0	1	2	0	1	0	0	0	0	0	0	0	0	0
FR	20	1	2	13	1	2	7	0	0	0	0	0	0	0	0
GER	25	9	1	0	0	0	10	2	0	15	7	1	0	0	0
IS	1	6	0	0	0	0	0	0	0	0	0	0	1	6	0
OO	24	2	24	14	2	12	8	0	0	2	0	0	0	0	12
PR	2	0	1	2	0	1	0	0	0	0	0	0	0	0	0
RUSS	2	0	3	2	0	1	0	0	2	0	0	0	0	0	0
SOV	2	0	2	0	0	0	0	0	2	2	0	0	0	0	0
USA	44	3	14	16	0	1	6	0	3	22	3	10	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.72	0.05	0.23	0.94	0.00	0.06	0.67	0.00	0.33	0.63	0.09	0.29	na	na	na
BR	0.85	0.03	0.12	0.93	0.00	0.07	0.92	0.08	0.00	0.57	0.00	0.43	na	na	na
GER	0.71	0.24	0.05	0.67	0.00	0.33	0.83	0.17	0.00	0.65	0.30	0.04	na	na	na
IS	0.14	0.86	0.00	na	na	na	na	na	na	na	na	na	0.14	0.86	0.00

C. "AEROA"

AIR FORCE ADVANTAGE ATTACKER WINS

	OVERALL			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D
AUS	0	0	0	0	0	0	0	0	0
BR	4	0	3	4	0	3	0	0	0
CS	0	0	0	0	0	0	0	0	0
ENG	0	0	0	0	0	0	0	0	0
FR	0	0	0	0	0	0	0	0	0
GER	15	7	1	15	7	1	0	0	0
IS	1	6	0	0	0	0	1	6	0
OO	2	0	12	2	0	0	0	0	12
PR	0	0	0	0	0	0	0	0	0
RUSS	0	0	0	0	0	0	0	0	0
SOV	2	0	0	2	0	0	0	0	0
USA	22	3	10	22	3	10	0	0	0

	OVERALL			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D
USA	0.63	0.09	0.29	0.63	0.09	0.29	na	na	na
BR	0.57	0.00	0.43	0.57	0.00	0.43	na	na	na
GER	0.65	0.30	0.04	0.65	0.30	0.04	na	na	na
IS	na	na	na	na	na	na	0.14	0.86	0.00

AIR FORCE ADVANTAGE ATTACKER LOSES

	OVERALL			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D
AUS	0	0	0	0	0	0	0	0	0
BR	0	7	0	0	7	0	0	0	0
CS	0	0	0	0	0	0	0	0	0
ENG	0	0	0	0	0	0	0	0	0
FR	0	0	0	0	0	0	0	0	0
GER	3	1	19	3	1	19	0	0	0
IS	1	6	0	0	0	0	1	6	0
OO	3	0	11	0	0	2	3	0	9
PR	0	0	0	0	0	0	0	0	0
RUSS	0	0	0	0	0	0	0	0	0
SOV	1	1	0	1	1	0	0	0	0
USA	0	34	1	0	34	1	0	0	0

	OVERALL			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D
USA	0.00	0.97	0.03	0.00	0.97	0.03	na	na	na
BR	0.00	1.00	0.00	0.00	1.00	0.00	na	na	na
GER	0.13	0.04	0.83	0.13	0.04	0.83	na	na	na
IS	na	na	na	na	na	na	0.14	0.86	0.00

D. "LEADA"

LEADERSHIP ADVANTAGE ATTACKER WINS

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	6	4	0	3	4	0	3	0	0	0	0	0	0	0	0
BR	28	8	1	4	4	1	7	4	0	17	0	0	0	0	0
CS	0	5	0	0	5	0	0	0	0	0	0	0	0	0	0
ENG	2	5	0	2	5	0	0	0	0	0	0	0	0	0	0
FR	14	22	0	10	21	0	4	1	0	0	0	0	0	0	0
GER	27	15	0	5	3	0	8	6	0	14	6	0	0	0	0
IS	1	28	0	0	0	0	0	0	0	0	0	0	1	28	0
OO	16	39	7	12	33	2	4	5	0	0	1	0	0	0	5
PR	1	10	0	1	10	0	0	0	0	0	0	0	0	0	0
RUSS	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0
SOV	14	6	2	0	0	0	1	0	0	13	6	2	0	0	0
USA	96	12	2	8	12	0	29	0	2	59	0	0	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.87	0.11	0.02	0.40	0.60	0.00	0.94	0.00	0.06	1.00	0.00	0.00	na	na	na
BR	0.68	0.30	0.02	0.38	0.56	0.06	0.64	0.36	0.00	1.00	0.00	0.00	na	na	na
GER	0.53	0.47	0.00	0.32	0.68	0.00	0.57	0.43	0.00	0.70	0.30	0.00	na	na	na
IS	0.03	0.97	0.00	na	na	na	na	na	na	na	na	na	0.03	0.97	0.00

LEADERSHIP ADVANTAGE ATTACKER LOSES

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	6	0	6	3	0	5	3	0	1	0	0	0	0	0	0
BR	22	1	8	3	1	7	12	0	1	7	0	0	0	0	0
CS	12	1	4	12	1	4	0	0	0	0	0	0	0	0	0
ENG	1	0	2	1	0	2	0	0	0	0	0	0	0	0	0
FR	13	1	9	7	1	8	6	0	1	0	0	0	0	0	0
GER	30	2	3	0	0	0	10	0	2	20	2	1	0	0	0
IS	3	4	0	0	0	0	0	0	0	0	0	0	3	4	0
OO	17	2	31	7	2	19	8	0	0	2	0	0	0	0	12
PR	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
RUSS	2	1	2	1	1	1	1	0	1	0	0	0	0	0	0
SOV	2	0	2	0	0	0	0	0	2	2	0	0	0	0	0
USA	42	0	19	8	0	9	5	0	4	29	0	6	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.69	0.00	0.31	0.47	0.00	0.53	0.56	0.00	0.44	0.83	0.00	0.17	na	na	na
BR	0.68	0.03	0.29	0.29	0.07	0.64	0.92	0.00	0.08	1.00	0.00	0.00	na	na	na
GER	0.82	0.08	0.11	0.33	0.33	0.33	0.83	0.00	0.17	0.87	0.09	0.04	na	na	na
IS	0.43	0.57	0.00	na	na	na	na	na	na	na	na	na	0.43	0.57	0.00

E. "TRNGA"

TRAINING ADVANTAGE ATTACKER WINS

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	8	2	0	5	2	0	3	0	0	0	0	0	0	0	0
BR	21	7	9	3	6	0	10	1	0	8	0	9	0	0	0
CS	5	0	0	5	0	0	0	0	0	0	0	0	0	0	0
ENG	4	3	0	4	3	0	0	0	0	0	0	0	0	0	0
FR	25	4	7	22	4	5	3	0	2	0	0	0	0	0	0
GER	24	17	1	6	2	0	10	4	0	8	11	1	0	0	0
IS	0	29	0	0	0	0	0	0	0	0	0	0	0	29	0
OO	54	5	3	42	3	2	8	1	0	0	1	0	4	0	1
PR	10	1	0	10	1	0	0	0	0	0	0	0	0	0	0
RUSS	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0
SOV	11	6	5	0	0	0	1	0	0	10	6	5	0	0	0
USA	70	13	27	12	8	0	7	0	24	51	5	3	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.64	0.12	0.25	0.60	0.40	0.00	0.23	0.00	0.77	0.86	0.08	0.05	na	na	na
BR	0.57	0.23	0.20	0.44	0.56	0.00	0.91	0.09	0.00	0.47	0.00	0.53	na	na	na
GER	0.64	0.34	0.02	0.84	0.16	0.00	0.71	0.29	0.00	0.40	0.55	0.05	na	na	na
IS	0.00	1.00	0.00	na	na	na	na	na	na	na	na	na	0.00	1.00	0.00

TRAINING ADVANTAGE ATTACKER LOSES

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	11	0	1	7	0	1	4	0	0	0	0	0	0	0	0
BR	24	4	3	7	4	0	13	0	0	4	0	3	0	0	0
CS	17	0	0	17	0	0	0	0	0	0	0	0	0	0	0
ENG	2	0	1	2	0	1	0	0	0	0	0	0	0	0	0
FR	19	1	3	12	1	3	7	0	0	0	0	0	0	0	0
GER	25	10	0	0	0	0	10	2	0	15	8	0	0	0	0
IS	3	4	0	0	0	0	0	0	0	0	0	0	3	4	0
OO	26	3	22	16	3	9	7	0	1	2	0	0	0	0	12
PR	2	0	1	2	0	1	0	0	0	0	0	0	0	0	0
RUSS	3	0	2	2	0	1	1	0	1	0	0	0	0	0	0
SOV	4	0	0	0	0	0	2	0	0	2	0	0	0	0	0
USA	42	2	17	16	0	1	0	0	9	26	2	7	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.69	0.03	0.28	0.94	0.00	0.06	0.00	0.00	1.00	0.74	0.06	0.20	na	na	na
BR	0.76	0.12	0.12	0.64	0.29	0.07	1.00	0.00	0.00	0.57	0.00	0.43	na	na	na
GER	0.71	0.26	0.03	0.67	0.00	0.33	0.83	0.17	0.00	0.65	0.35	0.00	na	na	na
IS	0.43	0.57	0.00	na	na	na	na	na	na	na	na	na	0.43	0.57	0.00

F. "MORALA"

MORAL ADVANTAGE ATTACKER WINS

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	10	0	0	7	0	0	3	0	0	0	0	0	0	0	0
BR	32	5	0	8	1	0	7	4	0	17	0	0	0	0	0
CS	4	1	0	4	1	0	0	0	0	0	0	0	0	0	0
ENG	7	0	0	7	0	0	0	0	0	0	0	0	0	0	0
FR	26	9	1	25	5	1	1	4	0	0	0	0	0	0	0
GER	35	7	0	8	0	0	10	4	0	17	3	0	0	0	0
IS	19	10	0	0	0	0	0	0	0	0	0	0	19	10	0
OO	53	9	0	41	6	0	9	0	0	0	1	0	3	2	0
PR	11	0	0	11	0	0	0	0	0	0	0	0	0	0	0
RUSS	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0
SOV	2	19	1	0	0	0	1	0	0	1	19	1	0	0	0
USA	60	50	0	15	5	0	0	31	0	45	14	0	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.55	0.45	0.00	0.75	0.25	0.00	0.00	1.00	0.00	0.76	0.24	0.00	na	na	na
BR	0.89	0.11	0.00	0.94	0.06	0.00	0.64	0.36	0.00	1.00	0.00	0.00	na	na	na
GER	0.87	0.13	0.00	1.00	0.00	0.00	0.71	0.29	0.00	0.85	0.15	0.00	na	na	na
IS	0.66	0.34	0.00	na	na	na	na	na	na	na	na	na	0.66	0.34	0.00

MORAL ADVANTAGE ATTACKER LOSES

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	11	1	0	7	1	0	4	0	0	0	0	0	0	0	0
BR	30	1	0	10	1	0	13	0	0	7	0	0	0	0	0
CS	15	0	2	15	0	2	0	0	0	0	0	0	0	0	0
ENG	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0
FR	21	1	1	14	1	1	7	0	0	0	0	0	0	0	0
GER	32	1	2	0	0	0	11	0	1	21	1	1	0	0	0
IS	4	2	1	0	0	0	0	0	0	0	0	0	4	2	1
OO	49	0	2	27	0	1	8	0	0	2	0	0	11	0	1
PR	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0
RUSS	5	0	0	3	0	0	2	0	0	0	0	0	0	0	0
SOV	2	2	0	0	0	0	2	0	0	0	2	0	0	0	0
USA	47	14	0	16	1	0	0	9	0	31	4	0	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.77	0.23	0.00	0.94	0.06	0.00	0.00	1.00	0.00	0.89	0.11	0.00	na	na	na
BR	0.97	0.03	0.00	0.93	0.07	0.00	1.00	0.00	0.00	1.00	0.00	0.00	na	na	na
GER	0.92	0.03	0.05	1.00	0.00	0.00	0.92	0.00	0.08	0.91	0.04	0.04	na	na	na
IS	0.57	0.29	0.14	na	na	na	na	na	na	na	na	na	0.57	0.29	0.14

G. "LOGSA"

LOGISTICS ADVANTAGE ATTACKER WINS

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	9	1	0	6	1	0	3	0	0	0	0	0	0	0	0
BR	31	4	2	9	0	0	10	0	1	12	4	1	0	0	0
CS	5	0	0	5	0	0	0	0	0	0	0	0	0	0	0
ENG	7	0	0	7	0	0	0	0	0	0	0	0	0	0	0
FR	36	0	0	31	0	0	5	0	0	0	0	0	0	0	0
GER	34	5	3	7	1	0	9	3	2	18	1	1	0	0	0
IS	29	0	0	0	0	0	0	0	0	0	0	0	29	0	0
OO	57	5	0	46	1	0	5	4	0	1	0	0	5	0	0
PR	11	0	0	11	0	0	0	0	0	0	0	0	0	0	0
RUSS	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0
SOV	7	15	0	0	0	0	0	1	0	7	14	0	0	0	0
USA	93	15	2	19	1	0	31	0	0	43	14	2	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.85	0.14	0.02	0.95	0.05	0.00	1.00	0.00	0.00	0.73	0.24	0.03	na	na	na
BR	0.86	0.09	0.05	1.00	0.00	0.00	0.91	0.00	0.09	0.71	0.24	0.06	na	na	na
GER	0.85	0.09	0.06	0.95	0.05	0.00	0.64	0.21	0.14	0.90	0.05	0.05	na	na	na
IS	1.00	0.00	0.00	na	na	na	na	na	na	na	na	na	1.00	0.00	0.00

LOGISTICS ADVANTAGE ATTACKER LOSES

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	11	0	1	8	0	0	3	0	1	0	0	0	0	0	0
BR	26	2	3	9	1	1	11	0	2	6	1	0	0	0	0
CS	16	0	1	16	0	1	0	0	0	0	0	0	0	0	0
ENG	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0
FR	21	0	2	14	0	2	7	0	0	0	0	0	0	0	0
GER	30	1	4	0	0	0	11	0	1	19	1	3	0	0	0
IS	7	0	0	0	0	0	0	0	0	0	0	0	7	0	0
OO	49	1	1	27	1	0	7	0	1	2	0	0	12	0	0
PR	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0
RUSS	4	0	1	2	0	1	2	0	0	0	0	0	0	0	0
SOV	4	0	0	0	0	0	2	0	0	2	0	0	0	0	0
USA	56	2	3	17	0	0	9	0	0	30	2	3	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.92	0.03	0.05	1.00	0.00	0.00	1.00	0.00	0.00	0.86	0.06	0.09	na	na	na
BR	0.85	0.06	0.09	0.86	0.07	0.07	0.85	0.00	0.15	0.86	0.14	0.00	na	na	na
GER	0.87	0.03	0.11	1.00	0.00	0.00	0.92	0.00	0.08	0.83	0.04	0.13	na	na	na
IS	1.00	0.00	0.00	na	na	na	na	na	na	na	na	na	1.00	0.00	0.00

H. "MOMNTA"

MOMENTUM ADVANTAGE ATTACKER WINS

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	9	1	0	6	1	0	3	0	0	0	0	0	0	0	0
BR	28	9	0	7	2	0	6	5	0	15	2	0	0	0	0
CS	3	2	0	3	2	0	0	0	0	0	0	0	0	0	0
ENG	7	0	0	7	0	0	0	0	0	0	0	0	0	0	0
FR	29	7	0	25	6	0	4	1	0	0	0	0	0	0	0
GER	23	19	0	5	3	0	11	3	0	7	13	0	0	0	0
IS	16	13	0	0	0	0	0	0	0	0	0	0	16	13	0
OO	48	13	1	37	9	1	8	1	0	0	1	0	3	2	0
PR	11	0	0	11	0	0	0	0	0	0	0	0	0	0	0
RUSS	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0
SOV	6	16	0	0	0	0	1	0	0	5	16	0	0	0	0
USA	74	36	0	14	6	0	26	5	0	34	25	0	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.67	0.33	0.00	0.70	0.30	0.00	0.84	0.16	0.00	0.58	0.42	0.00	na	na	na
BR	0.80	0.20	0.00	0.88	0.13	0.00	0.55	0.45	0.00	0.88	0.12	0.00	na	na	na
GER	0.64	0.36	0.00	0.84	0.16	0.00	0.79	0.21	0.00	0.35	0.65	0.00	na	na	na
IS	0.55	0.45	0.00	na	na	na	na	na	na	na	na	na	0.55	0.45	0.00

MOMENTUM ADVANTAGE ATTACKER LOSES

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	12	0	0	8	0	0	4	0	0	0	0	0	0	0	0
BR	26	5	0	10	1	0	9	4	0	7	0	0	0	0	0
CS	16	1	0	16	1	0	0	0	0	0	0	0	0	0	0
ENG	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0
FR	19	4	0	13	3	0	6	1	0	0	0	0	0	0	0
GER	26	9	0	0	0	0	8	4	0	18	5	0	0	0	0
IS	4	2	1	0	0	0	0	0	0	0	0	0	4	2	1
OO	48	3	0	27	1	0	8	0	0	2	0	0	11	1	0
PR	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0
RUSS	5	0	0	3	0	0	2	0	0	0	0	0	0	0	0
SOV	2	2	0	0	0	0	2	0	0	0	2	0	0	0	0
USA	54	5	2	16	1	0	9	0	0	29	4	2	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.89	0.08	0.03	0.94	0.06	0.00	1.00	0.00	0.00	0.83	0.11	0.06	na	na	na
BR	0.85	0.15	0.00	0.93	0.07	0.00	0.69	0.31	0.00	1.00	0.00	0.00	na	na	na
GER	0.76	0.24	0.00	1.00	0.00	0.00	0.67	0.33	0.00	0.78	0.22	0.00	na	na	na
IS	0.57	0.29	0.14	na	na	na	na	na	na	na	na	na	0.57	0.29	0.14

I. "INTELA"

INTELLIGENCE ADVANTAGE ATTACKER WINS

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	6	4	0	3	4	0	3	0	0	0	0	0	0	0	0
BR	30	5	2	6	1	2	9	2	0	15	2	0	0	0	0
CS	3	2	0	3	2	0	0	0	0	0	0	0	0	0	0
ENG	6	1	0	6	1	0	0	0	0	0	0	0	0	0	0
FR	27	9	0	22	9	0	5	0	0	0	0	0	0	0	0
GER	24	14	4	7	1	0	9	5	0	8	8	4	0	0	0
IS	27	2	0	0	0	0	0	0	0	0	0	0	27	2	0
OO	46	16	0	35	12	0	7	2	0	1	0	0	3	2	0
PR	8	3	0	8	3	0	0	0	0	0	0	0	0	0	0
RUSS	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0
SOV	9	13	0	0	0	0	0	1	0	9	12	0	0	0	0
USA	104	4	2	15	4	1	31	0	0	58	0	1	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.95	0.04	0.02	0.75	0.20	0.05	1.00	0.00	0.00	0.98	0.00	0.02	na	na	na
BR	0.82	0.14	0.05	0.75	0.13	0.13	0.82	0.18	0.00	0.88	0.12	0.00	na	na	na
GER	0.60	0.32	0.08	0.79	0.21	0.00	0.64	0.36	0.00	0.40	0.40	0.20	na	na	na
IS	0.93	0.07	0.00	na	na	na	na	na	na	na	na	na	0.93	0.07	0.00

INTELLIGENCE ADVANTAGE ATTACKER LOSES

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	8	1	3	6	1	1	2	0	2	0	0	0	0	0	0
BR	26	0	5	7	0	4	13	0	0	6	0	1	0	0	0
CS	14	0	3	14	0	3	0	0	0	0	0	0	0	0	0
ENG	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0
FR	18	1	4	12	1	3	6	0	1	0	0	0	0	0	0
GER	28	2	5	0	0	0	9	1	2	19	1	3	0	0	0
IS	6	0	1	0	0	0	0	0	0	0	0	0	6	0	1
OO	43	2	6	21	2	5	8	0	0	2	0	0	11	0	1
PR	2	0	1	2	0	1	0	0	0	0	0	0	0	0	0
RUSS	4	0	1	3	0	0	1	0	1	0	0	0	0	0	0
SOV	4	0	0	0	0	0	2	0	0	2	0	0	0	0	0
USA	50	1	10	13	1	3	9	0	0	28	0	7	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.82	0.02	0.16	0.76	0.06	0.18	1.00	0.00	0.00	0.80	0.00	0.20	na	na	na
BR	0.85	0.00	0.15	0.71	0.00	0.29	1.00	0.00	0.00	0.86	0.00	0.14	na	na	na
GER	0.79	0.05	0.16	0.67	0.00	0.33	0.75	0.08	0.17	0.83	0.04	0.13	na	na	na
IS	0.86	0.00	0.14	na	na	na	na	na	na	na	na	na	0.86	0.00	0.14

J. "TECHNA"

TECHNOLOGY ADVANTAGE ATTACKER WINS

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	6	4	0	3	4	0	3	0	0	0	0	0	0	0	0
BR	24	13	0	7	2	0	5	6	0	12	5	0	0	0	0
CS	3	2	0	3	2	0	0	0	0	0	0	0	0	0	0
ENG	5	2	0	5	2	0	0	0	0	0	0	0	0	0	0
FR	25	11	0	21	10	0	4	1	0	0	0	0	0	0	0
GER	22	20	0	7	1	0	6	8	0	9	11	0	0	0	0
IS	16	13	0	0	0	0	0	0	0	0	0	0	16	13	0
OO	36	26	0	32	15	0	4	5	0	0	1	0	0	5	0
PR	8	3	0	8	3	0	0	0	0	0	0	0	0	0	0
RUSS	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0
SOV	16	6	0	0	0	0	0	1	0	16	5	0	0	0	0
USA	88	22	0	16	4	0	19	12	0	53	6	0	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.80	0.20	0.00	0.80	0.20	0.00	0.61	0.39	0.00	0.90	0.10	0.00	na	na	na
BR	0.66	0.34	0.00	0.75	0.25	0.00	0.45	0.55	0.00	0.71	0.29	0.00	na	na	na
GER	0.57	0.43	0.00	0.79	0.21	0.00	0.43	0.57	0.00	0.45	0.55	0.00	na	na	na
IS	0.55	0.45	0.00	na	na	na	na	na	na	na	na	na	0.55	0.45	0.00

TECHNOLOGY ADVANTAGE ATTACKER LOSES

	OVERALL			1600-1913			1913-1939			1939-1945			1945-2000		
RowNames	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	12	0	0	8	0	0	4	0	0	0	0	0	0	0	0
BR	30	1	0	11	0	0	12	1	0	7	0	0	0	0	0
CS	17	0	0	17	0	0	0	0	0	0	0	0	0	0	0
ENG	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0
FR	22	1	0	16	0	0	6	1	0	0	0	0	0	0	0
GER	33	2	0	0	0	0	11	1	0	22	1	0	0	0	0
IS	7	0	0	0	0	0	0	0	0	0	0	0	7	0	0
OO	46	0	5	25	0	3	8	0	0	0	0	2	12	0	0
PR	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0
RUSS	5	0	0	3	0	0	2	0	0	0	0	0	0	0	0
SOV	4	0	0	0	0	0	2	0	0	2	0	0	0	0	0
USA	57	4	0	16	1	0	9	0	0	32	3	0	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.93	0.07	0.00	0.94	0.06	0.00	1.00	0.00	0.00	0.91	0.09	0.00	na	na	na
BR	0.97	0.03	0.00	1.00	0.00	0.00	0.92	0.08	0.00	1.00	0.00	0.00	na	na	na
GER	0.95	0.05	0.00	1.00	0.00	0.00	0.92	0.08	0.00	0.96	0.04	0.00	na	na	na
IS	1.00	0.00	0.00	na	na	na	na	na	na	na	na	na	1.00	0.00	0.00

K. "INITA"

INITIATIVE ADVANTAGE ATTACKER WINS

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	6	4	0	4	3	0	2	1	0	0	0	0	0	0	0
BR	8	29	0	3	6	0	2	9	0	3	14	0	0	0	0
CS	1	4	0	1	4	0	0	0	0	0	0	0	0	0	0
ENG	1	6	0	1	6	0	0	0	0	0	0	0	0	0	0
FR	3	33	0	3	28	0	0	5	0	0	0	0	0	0	0
GER	6	36	0	1	7	0	5	9	0	0	20	0	0	0	0
IS	3	26	0	0	0	0	0	0	0	0	0	0	3	26	0
OO	13	49	0	6	41	0	2	7	0	0	1	0	5	0	0
PR	2	9	0	2	9	0	0	0	0	0	0	0	0	0	0
RUSS	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0
SOV	4	18	0	0	0	0	0	1	0	4	17	0	0	0	0
USA	15	95	0	1	19	0	1	30	0	13	46	0	0	0	0

	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.14	0.86	0.00	0.05	0.95	0.00	0.03	0.97	0.00	0.22	0.78	0.00	na	na	na
BR	0.20	0.80	0.00	0.25	0.75	0.00	0.18	0.82	0.00	0.18	0.82	0.00	na	na	na
GER	0.15	0.85	0.00	0.16	0.84	0.00	0.36	0.64	0.00	0.00	1.00	0.00	na	na	na
IS	0.10	0.90	0.00	na	na	na	na	na	na	na	na	na	0.10	0.90	0.00

INITIATIVE ADVANTAGE ATTACKER LOSES

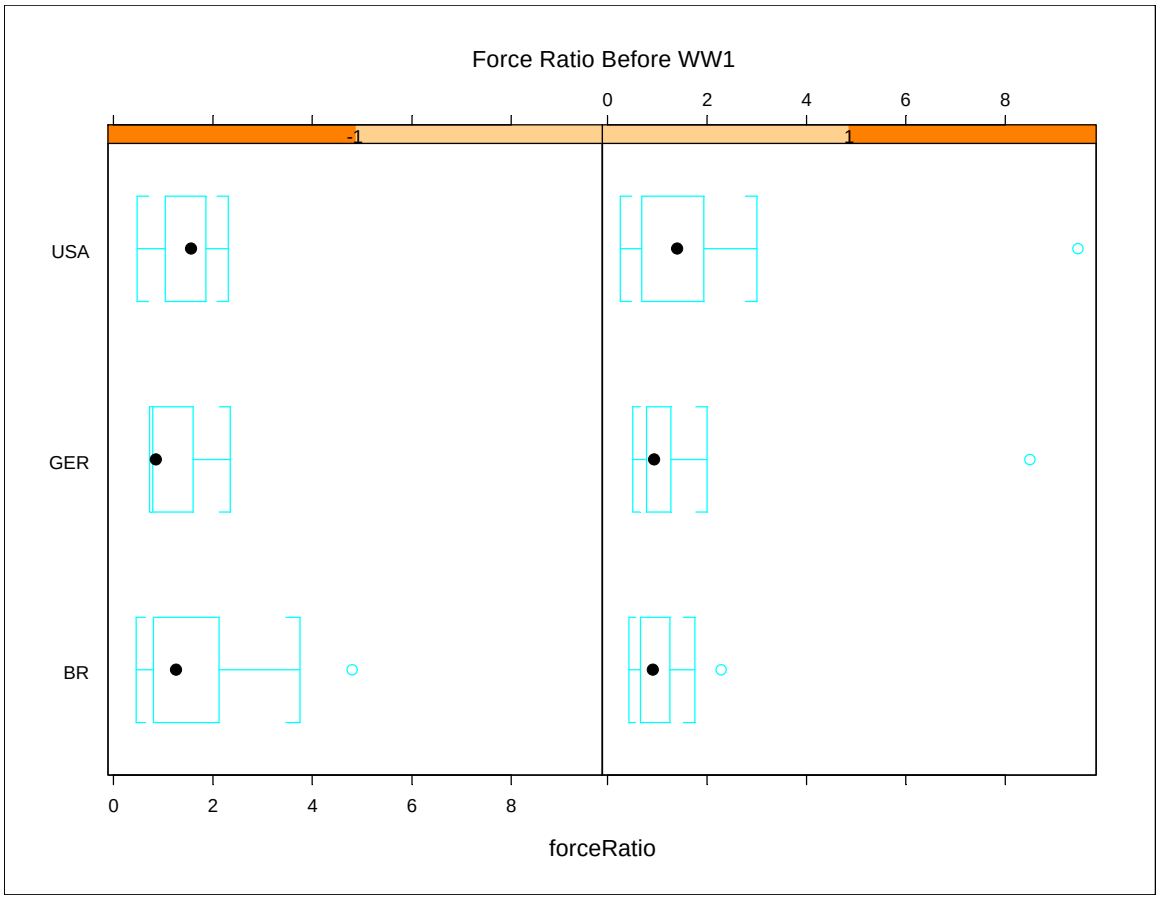
	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
AUS	9	1	2	6	1	1	3	0	1	0	0	0	0	0	0
BR	18	10	3	6	3	2	7	5	1	5	2	0	0	0	0
CS	3	14	0	3	14	0	0	0	0	0	0	0	0	0	0
ENG	0	3	0	0	3	0	0	0	0	0	0	0	0	0	0
FR	14	7	2	8	7	1	6	0	1	0	0	0	0	0	0
GER	12	20	3	0	0	0	6	5	1	6	15	2	0	0	0
IS	2	5	0	0	0	0	0	0	0	0	0	0	2	5	0
OO	32	13	5	17	8	3	6	2	0	1	1	0	8	2	2
PR	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0
RUSS	3	1	1	2	1	0	1	0	1	0	0	0	0	0	0
SOV	3	1	0	0	0	0	2	0	0	1	1	0	0	0	0
USA	24	29	8	4	11	2	4	4	1	16	14	5	0	0	0

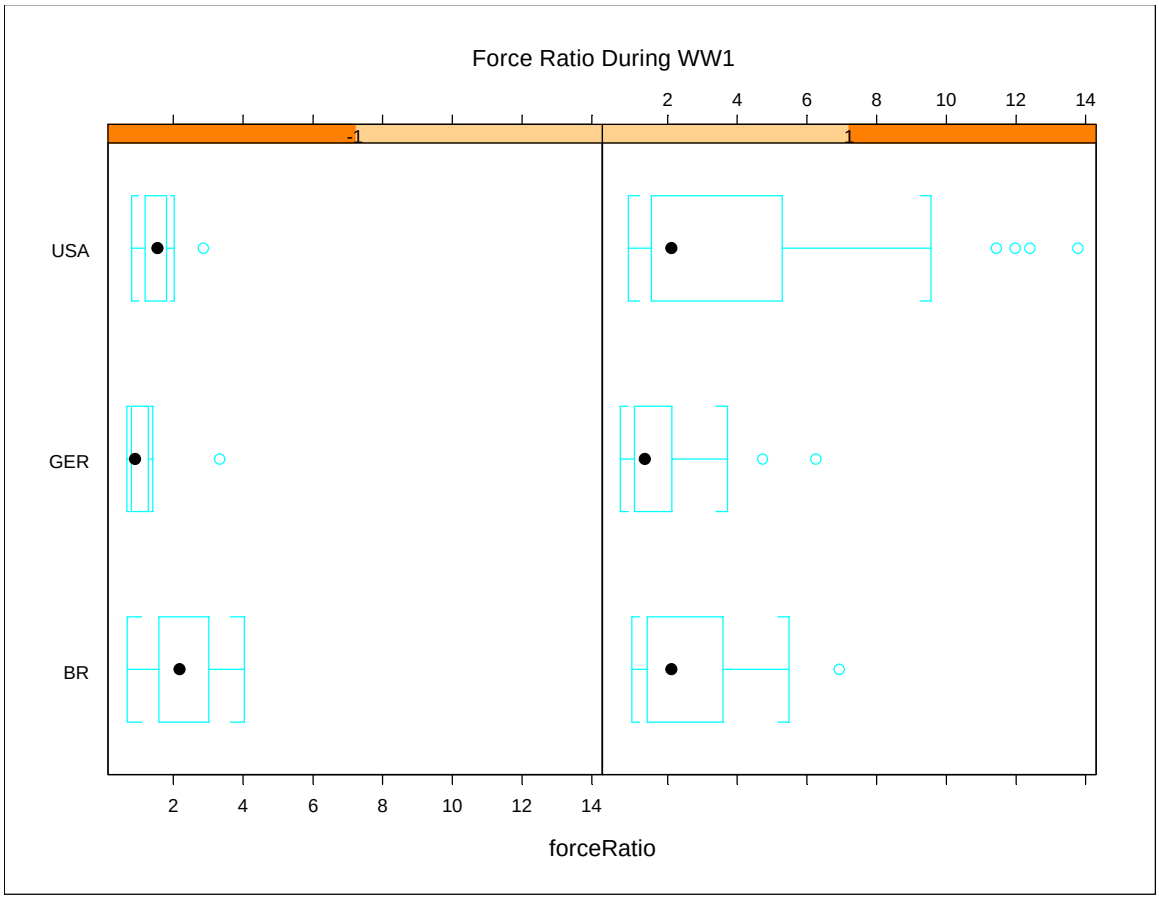
	OVERALL			1600-1914			1913-1939			1939-1945			1945-2000		
COUNTRY	O	A	D	O	A	D	O	A	D	O	A	D	O	A	D
USA	0.39	0.48	0.13	0.24	0.65	0.12	0.44	0.44	0.11	0.46	0.40	0.14	na	na	na
BR	0.53	0.38	0.09	0.43	0.43	0.14	0.54	0.38	0.08	0.71	0.29	0.00	na	na	na
GER	0.39	0.53	0.08	1.00	0.00	0.00	0.50	0.42	0.08	0.26	0.65	0.09	na	na	na
IS	0.29	0.71	0.00	na	na	na	na	na	na	na	na	na	0.29	0.71	0.00

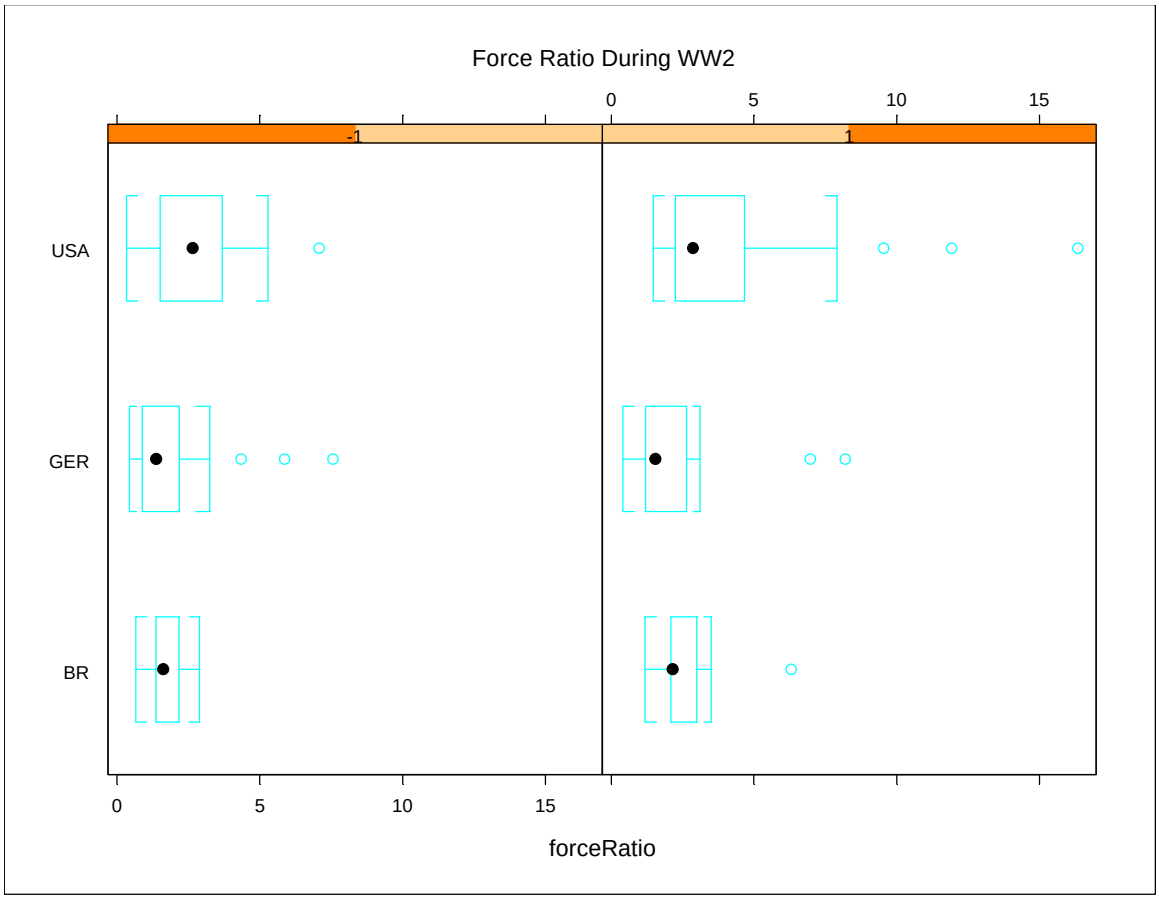
APPENDIX B. BOXPLOTS OF OBJECTIVE VARIABLES

This section has the boxplots that are not listed for the objective variables analyzed in the second chapter. The reader may refer to p.15 for more explanation on the boxplots.

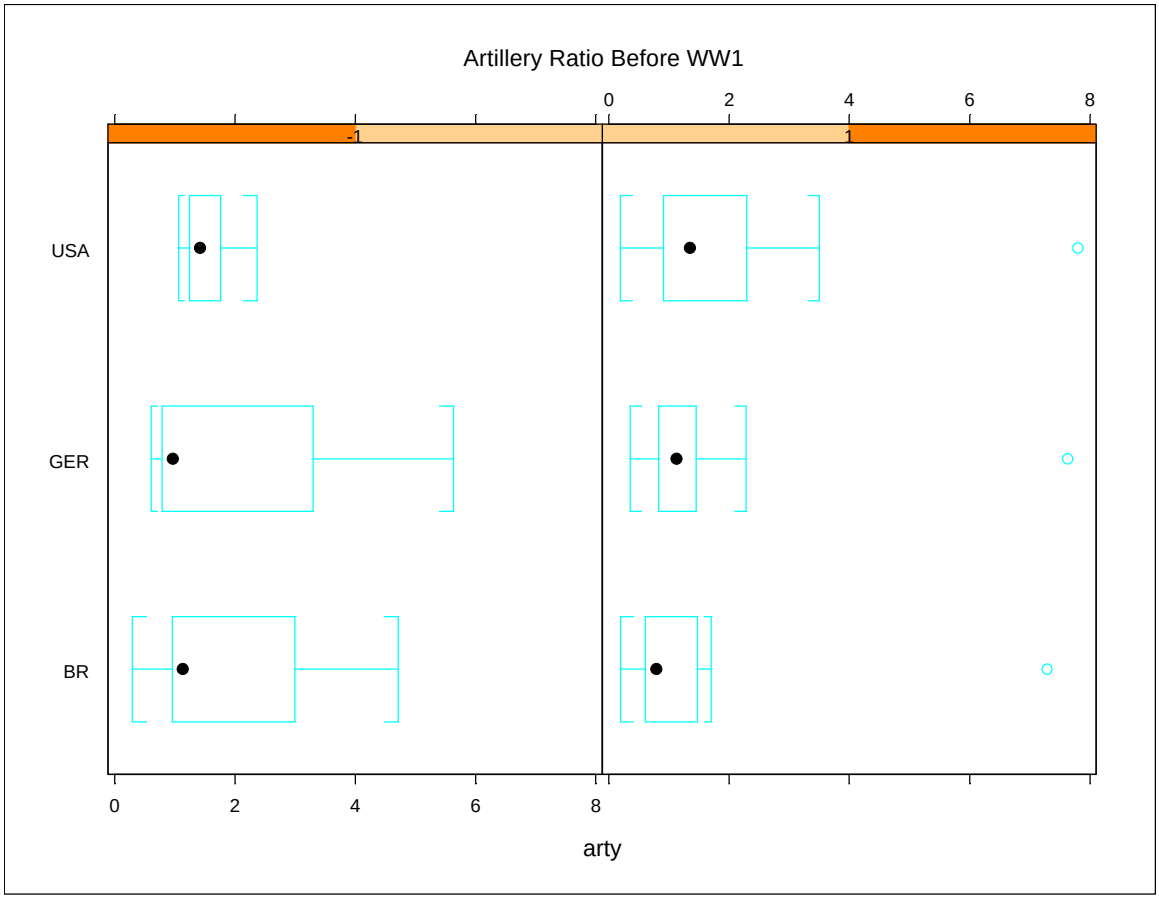
A. FORCE RATIO

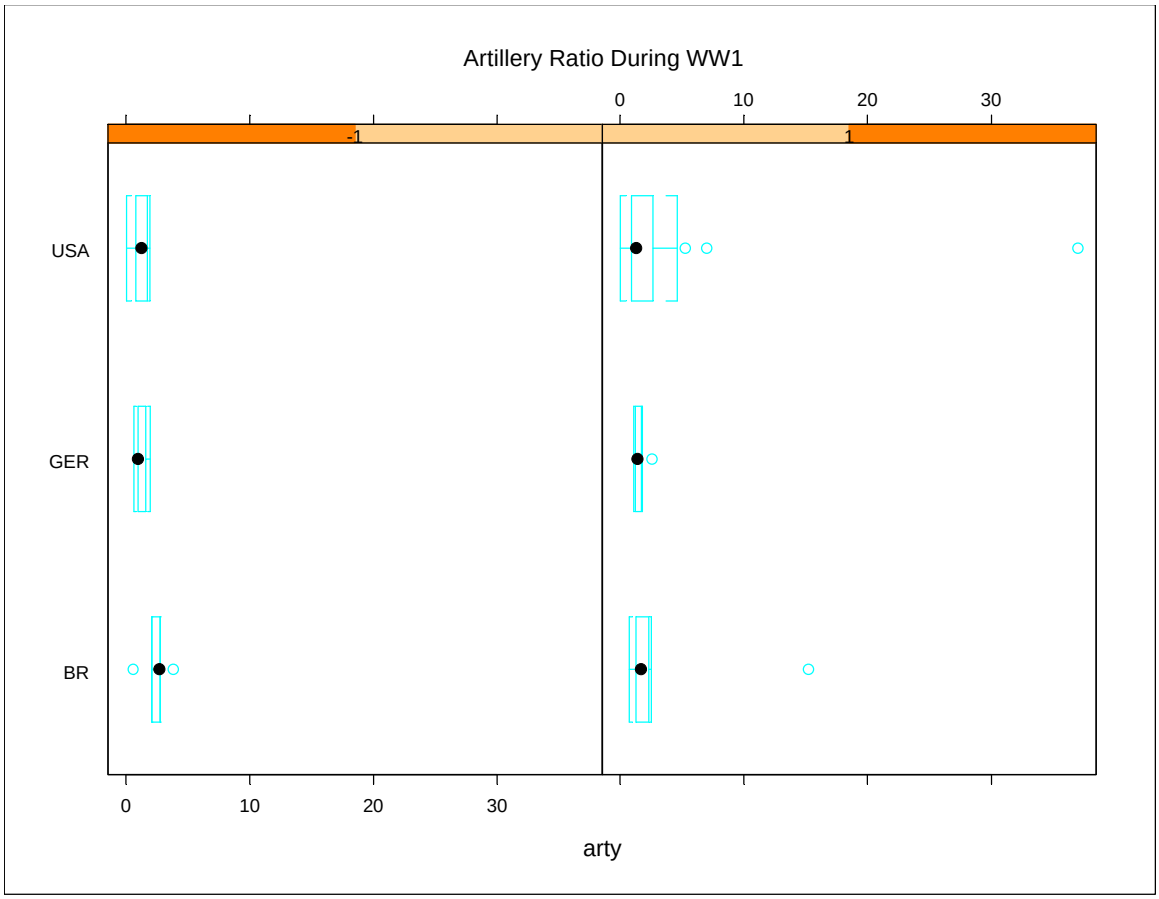


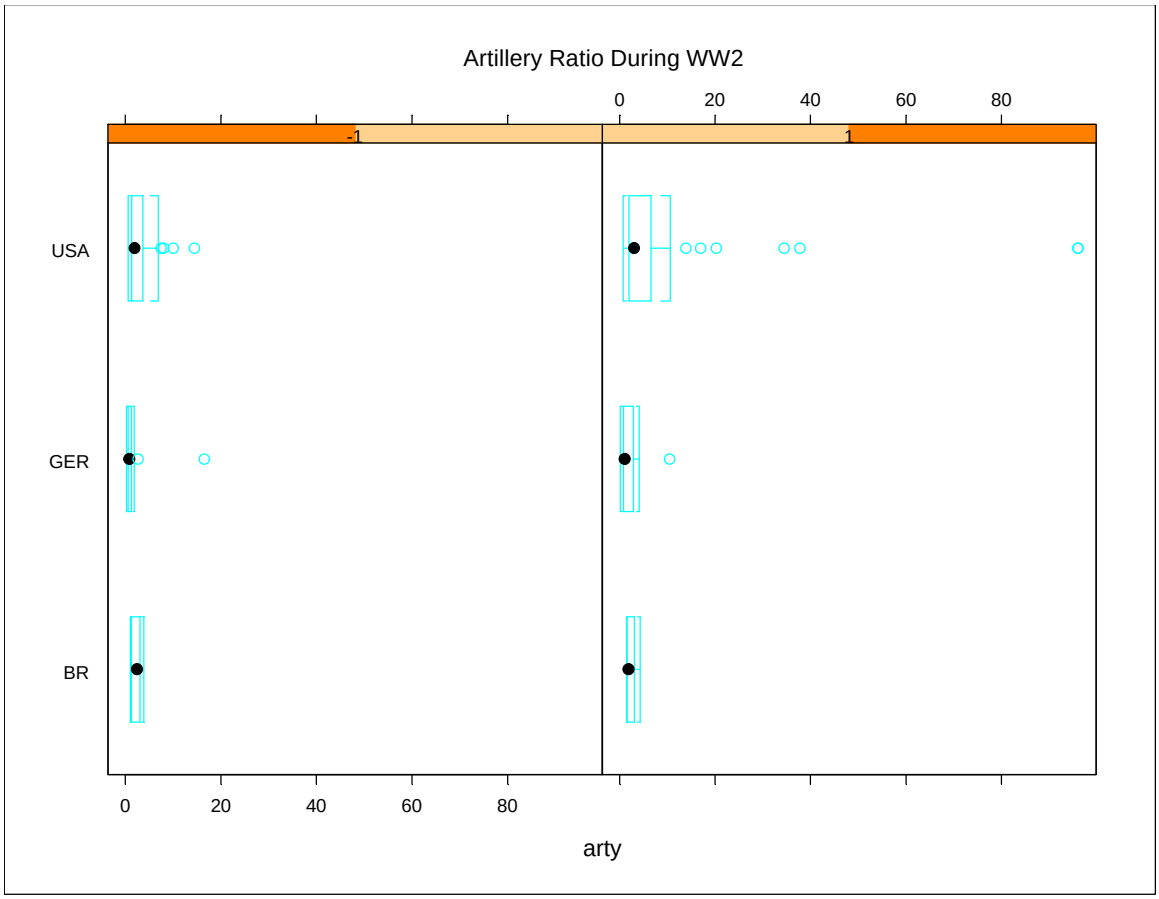

Force Ratio During WW1
2
4
6
8
10
12
14
-1
1
USA
GER
BR
2
4
6
8
10
12
14
forceRatio


Force Ratio During WW2
0
5
10
15
-1
1
USA
GER
BR
0
5
10
15
forceRatio

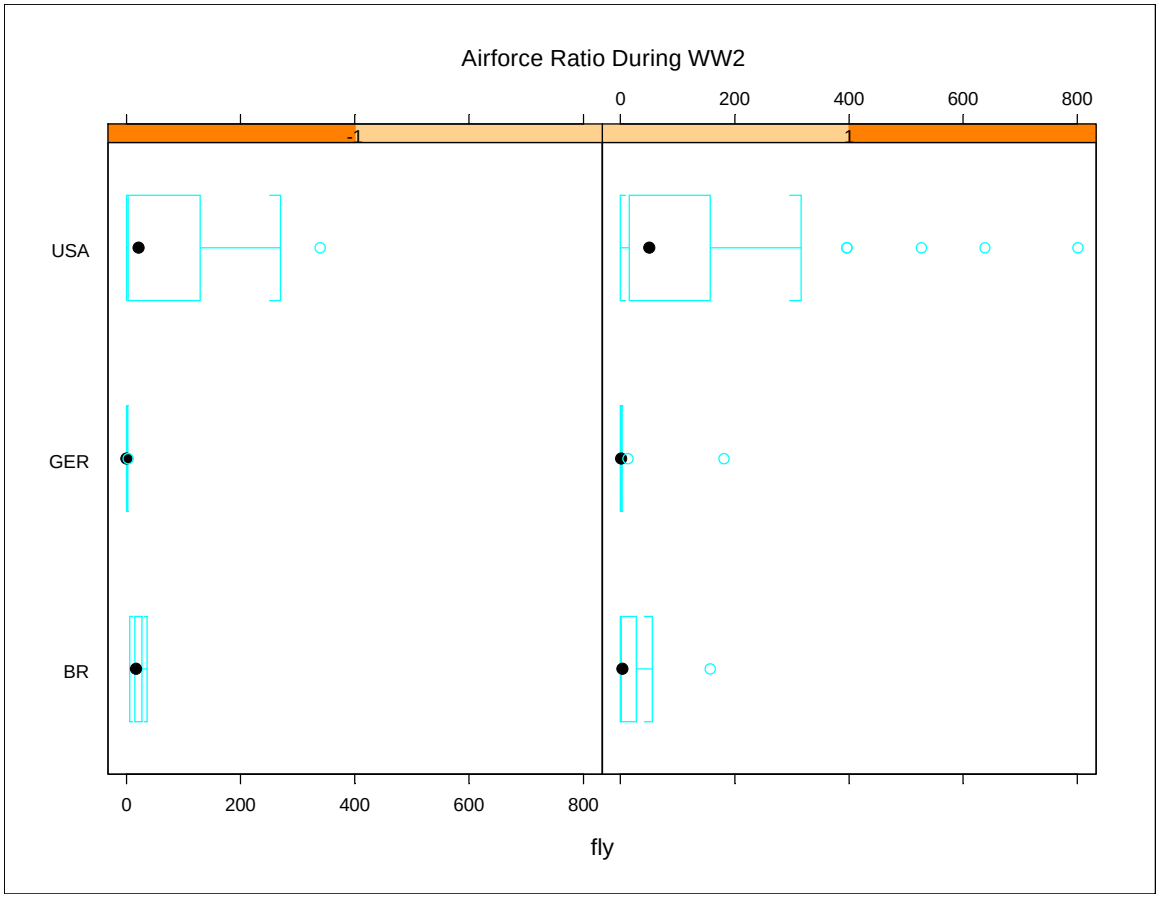
B. ARTILLERY RATIO




Artillery Ratio During WW1
0
10
20
30
-1
1
USA
GER
BR
0
10
20
30
arty


Artillery Ratio During WW2
0
20
40
60
80
-1
1
USA
GER
BR
0
20
40
60
80
arty

C. AIR FORCE RATIO



THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX C. ACRONYMS

COUNTRY NAMES

AUS:	Austria
ENG:	England
BR:	Britain
GER:	Germany
PRUSS:	Prussia
IS:	Israel
USA:	United States of America
SOV:	USSR
RUSS:	Russia
CS:	Confederate States (Present only in the battles of American Civil War)
TU:	Turkey
EG:	Egypt
SYR:	Syria

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

1. Hartley, Dean S., *"Topics in Operations Research: Predicting Combat Effects,"* Military Applications Society of INFORMS, 2001.

2. Dupuy, Col. T. N. (U. S. Army, Ret.), *"Numbers, Predictions and War,"* Hero Books, 1985.

3. Yigit, Faruk, *"Finding the Important Factors in Battle Outcomes: A Statistical Exploration of Data from Major Battles,"* Master's Thesis, Naval Postgraduate School, Monterey, California, 2000.

4. Coban, Muzaffer, *"Predicting Battle Outcomes with Classification Trees,"* Master's Thesis, Naval Postgraduate School, Monterey, California, 2001.

5. Personal Communication from Professor Thomas Lucas, Operational Research Department, Naval Postgraduate School, Monterey, California.

6. Slate Magazine, MSN, [http://slate.msn.com], February 26, 2003.

7. Chambers, John M. and Hastie, Trevor J., *"Statistical Models in S,"* Wadsworth & Brooks/Cole Advanced Books & Software, 1992.

8. Devore, Jay L., *"Probability and Statistics for Engineering and Sciences,"* Duxbury, 2000.

9. Venables, W. N. and Ripley, B. D., *"Modern Applied Statistics with S-PLUS,"* Springer, 1999.

10. Therneau, Terry M. and Atkinson, Elizabeth J., *"An Introduction to Recursive Partitioning Using the RPART Routines*," Mayo Foundation, September 3, 1997.

11. FM100-5 OPERATIONS, [http://usasma.bliss.army.mil/Pubs/FM_100-5/FM_100-5.pdf], May 8, 2003.

12. Class Notes OA 3103 Samuel E. Buttrey, Operational Research Department, Naval Postgraduate School, Monterey, California.

13 S-Plus 4, Guide to Statistics, Data Analysis Products Division MathSoft, Inc., Seattle, Washington.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Thomas W. Lucas
   Naval Postgraduate School
   Monterey, California

4. Samuel E. Buttrey
   Naval Postgraduate School
   Monterey, California

5. Ali Cakan
   Kara Kuvvetleri Personel Baskanligi
   Ankara, Turkey